The European Association for Digital Humanities (EADH)
Association for Computers and the Humanities (ACH)
Canadian Society for Digital Humanities / Société canadienne des humanités numériques (CSDH/SCHN)
centerNet
Australasian Association for Digital Humanities (aaDH)
Japanese Association for Digital Humanites (JADH)

Digital Humanities 2016

Conference Abstracts

Jagiellonian University
&
Pedagogical University

Kraków
11–16 July 2016

Kraków 2016

Edited by

Maciej Eder
Jan Rybicki

DHConvalidator service

Marco Petris

On-line abstracts

Michał Woźniak

Design and typesetting

Maciej Eder

Proof-reading

Aleksandra Ptasznik
Karolina Eder
Daria Wyka
Karolina Wróbel
Weronika Ślęczkowska
Anita Uryga
Sandra Romanowicz

ISBN 978–83–942760–3–4

Welcome to Digital Humanities 2016

The annual "Digital Humanities" conference, first held in 1989, gives a clear proof that the field of DH is, first and foremost, very well established, even if more and more subfields emerge almost every year. Secondly, the conference confirms the fact that the discipline is constantly growing. This year is no exception. Even more: the 27th joint conference of EADH (ALLC) and ACH, and the 8th conference under the auspices of ADHO, is by far the biggest event in the field, with its almost 450 accepted submissions in different categories: panels, long papers, short papers, posters, and pre-conference workshops. The number of registered participants exceeded 850 at the time of writing these words.

The conference takes place in Kraków; this is only the second time (after Debrecen 1998) that it comes to Central/Eastern Europe. The region's rich past and its recent rapid growth has inspired the conference theme, "Digital Identities: the Past and the Future". This theme aims at stressing the very strong connections between DH and its roots in the medieval idea of a university with a prominent role of the Liberal Arts. We strongly believe that the unique relation between the origins of the humanities' scholarship and the opportunities provided by computer algorithms and the enormous amount of resources (text collections, linguistic corpora, databases, virtual libraries) can lead to a new scientific revolution.

Kraków, the venue of the conference, has been a major center of learning and culture in this part of the world: the Jagiellonian, founded in 1364, is usually ranked first among Polish universities, and the same is true of the Kraków's Pedagogical University within Poland's quite extensive community of pedagogical universities. No wonder, then, that the conference is hosted jointly by those two institutions: the Jagiellonian University and the Pedagogical University of Kraków. Their collaboration is a manifestation of the vivid digital humanities scene emerging in Poland's major centre of learning and culture.

I would like to thank all those who submitted proposals this year and all those who agreed to act as reviewers. The work load for the group of scholars who undertook the demanding task of reviewing the submissions was significantly higher than in previous years – I would like to thank them all for their great job. I would like to give my thanks to the members of the Program Committee, who this year included: Diane Jakacki (CSDH/SCHN), Michael Eberle-Sinatra (CSDH/SCHN), Jennifer Guiliano (ACH), Brett D. Hirsch (aaDH), Leif Isaksen (EADH), Asanobu Kitamoto (JADH), Inna Kizhner (centerNet), Maurizio Lana (EADH), Kiyonori Nagasaki (JADH), Roopika Risam (ACH), Glenn Roe (aaDH), Sinai Rusinek (centerNet) and Deb Verhoeven (aaDH). My special and warmest thanks go to the Program Committee Chair Manfred Thaller, whose contribution to the conference was simply outstanding. I also want to mention the Local Organizers, who were brave enough to suggest Kraków as a potential host for the DH conference, and efficient enough to make it happen.

Karina van Dalen-Oskam
ADHO Steering Committee Chair

The 27th Joint International Conference of the Association for Literary
and Linguistic Computing and Association for Computers and the Humanities,
The 8th Joint International Conference of the Alliance of Digital Humanities Organizations

Program Committee

Manfred Thaller (chair)
Diane Jakacki (vice-chair)
Michael Eberle-Sinatra
Jennifer Guiliano
Brett D. Hirsch
Leif Isaksen
Asanobu Kitamoto
Inna Kizhner
Maurizio Lana
Kiyonori Nagasaki
Roopika Risam
Glenn Roe
Sinai Rusinek
Deb Verhoeven

Local Organizers

Maciej Eder
Jan Rybicki

Local Organizing Board

Katarzyna Bazarnik
Elżbieta Górska
Rafał Górski
Magda Heydel
Władysław Marek Kolasa
Krzysztof Nowak
Iwona Pietrzkiewicz
Maria Piotrowska
Bogusław Skowronek
Elżbieta Tabakowska
Ewa Willim
Grażyna Wrona

Gold Sponsor

GALE Cengage Learning

Organizers

Alliance of Digital Humanities Organizations (ADHO)
Jagiellonian University, Faculty of Philology
Pedagogical University in Kraków, Faculty of Philology

Table of Contents

Plenary lectures

Panels

Long papers

Short papers

Posters

Pre-conference workshops

Plenary lectures

Can CERN serve as a model for Digital Humanities?

Agnieszka Zalewska
Institute of Nuclear Physics, Polish Academy of Sciences

Nowadays CERN is perhaps the most recognized model of multinational collaboration in research. Created in 1954, it has given the international community some of the most fundamental insights into the essence of matter; and it has returned the investment in innumerable advances in technology. Its Large Hadron Collider is the world's largest particle accelerator; this is where the Higgs boson has been observed; and this is also where the Web was born to revolutionise the way we all communicate, as the DH community knows probably better than anyone.

CERN's success has always been attributed to its being a joint venture of an increasing number of nations. While it enjoys a stable status thanks to agreements between politicians, it would not have been possible without its visionaries; their vision extends from pure research to how people from all over the world and from a number of disciplines can work together and all contribute to that success.

Digital Humanities is equally and increasingly an international venture, as is evidenced, for one, by the map of past, present and future DH conferences. Here, too, people from all over the world, from various nations and ethnic groups, from very many scholarly disciplines want to work together. The lessons we have learned at CERN can help the DH community in terms of institutional organization and, more importantly, in the general approach to multinational and diverse scholarly collaboration.

Early Funding of Humanities Computing: A Personal History

Helen Agüera
National Endowment for the Humanities

From the 1970s to the 1990s, a time when funding opportunities for work in digital humanities were limited, the National Endowment for the Humanities (NEH) supported computer-based projects in several disciplines of the humanities. In this talk I review the agency's initial interest in the use of computer technology for research in the humanities. Through the lens of my experience as program officer for many of the early projects that employed computer methods, I discuss the transition from funding projects that used a computer as a tool to produce print publications, such as reference resources and scholarly editions, to supporting digital projects that published solely in electronic formats.

As digital projects became the norm in the humanities, the NEH extended its programs to support projects that address the needs of digital humanists, including development of digital standards, tools, and infrastructure. Besides describing these programs, I highlight NEH's collaboration with other funding organizations to facilitate work at the intersection of the humanities and the sciences. I conclude with some observations on the transformations that digital technology has brought to the conduct of collaborative research projects in the humanities and to the agency's evaluation process. (My presentation relates to my personal experience and does not represent the views of the NEH.)

Touching the interface: Bishop Cosin and unsolved problems in (digital) information design

Claire Warwick
Durham University

Some problems in the design of digital resources have turned out to be unexpectedly difficult to solve, for example: why is it difficult to locate ourselves and understand the extent and shape of digital information resources? Why is digital serendipity still so unusual? Why do users persist in making notes on paper rather than using digital annotation systems? Why do we like to visit and work in a library, and browse open stacks, even though we could access digital information remotely? Why do we still love printed books, but feel little affection for digital e-readers? Why are vinyl records so popular? Why is the experience of visiting a museum still relatively unaffected by digital interaction? The answer is very emphatically not because users are luddites, ill-informed, badly-trained or stupid.

I will argue that the reasons these problems persist may be due to the very complex relationship between physical and digital information, and information resources. I will discuss the importance of spatial orientation, memory, pleasure and multi-sensory input, especially touch, in making sense of, and connections between physical and digital information. I will also argue that, in this context, we have much to learn from the designers of early printed books and libraries, such John Cosin, a seventeenth-century bishop of Durham, who founded the little-known marvel that was the first public library in the North of England, and still exists, intact; one of the collections of Durham University library.

Panels

Playable Books at Electronic Literature's Interface

Katarzyna Bazarnik
k.bazarnik@uj.edu.pl
Jagellonian University, Korporacja Ha!art, Poland

Kathi Inman Berens
kathiberens@gmail.com
Portland State University, United States of America

Zenon Fajfer
zenkasi@wp.pl
Independent poet, Kraków, Poland

Susan Garfinkel
sgarfinkel@loc.gov
United States Library of Congress

For DH2016 with its theme of "Digital Identities: The Past and The Future" we propose a four-person panel in which we present the history of playable books (Garfinkel), the theory of playable books (Bazarnik), the artistic practice of playable books (Fajfer) and bookish aspects of digital literary games (Inman Berens). Together, we aim to demonstrate the procedural affinities between analog and digital modes of reading a literary interface.

This panel approaches the notion of play from the vantage of the physicality of books and what Johanna Drucker calls "performative materiality" (2013). When Espen Aarseth said in his 2015 keynote at the Electronic Literature Organization's conference that "games are the most important form of digital literature," it ruffled a few feathers. Here, we take Aarseth's provocation as a starting point from which we examine the past and future of what we call playable books. We will perform examples of digital and analog playable books, offering critical reflection toward developing a common language for methodological approaches across national languages and traditions, across analog and digital reading practices. The Digital Humanities 2016 conference is a unique opportunity to put global perspectives of ergodicity dynamically in dialog with each other. We expect, in other words, that our various types of expertise will co-mingle emergently as we "play" the works, isolate attributes, and foster conversation among the panelists and audience in this international venue. Liveness and co-presence are required to advance this form of cross-cultural communication.

All reading is situated and embodied. Throughout the history of the book, whether manuscript or mechanically reproduced, we see an ongoing tension in the moment of a reader's encounter with written work: between the abstract ideal of a text that stands on its own *sans* manifestations, and the actual instantiations of texts as books without whose affordance of interface the encounter would not be possible. Building on Aarseth (1997), Kirschenbaum (2008a; 2008b), Wardrip-Fruin (2010) and others, we note that the procedurality of playable books becomes newly and uniquely visible through the lens of computational materiality, that is, the turn to the digital. Books have always been random access portable storage devices. When Kirschenbaum argues that "new media cannot be studied apart from individual instances of inscription, object, and code as they propagate on, across, and through specific storage devices, operating systems, software environments and network protocols" (2008b: 23), we understand that such approaches prompt attention to the physicality of books. Jessica Pressman coined the term "bookishness" to describe "novels [that] exploit the power of the print page in ways that draw attention to the book as a multimedia format, one informed by and connected to digital technologies" (2009: 456). Zenon Fajfer (1999) describes "liberature" as literature which, in response to the digital media, foregrounds the shape and structure of the physical book as its semantically charged constituents. Hayles (2008) and Florian Cramer (2016) have separately written about the "post-digital" book, that is, printed books that manifest the aesthetic of Web and networked technologies, such as Mark Z. Danielewski's *House of Leaves* (2000).

As part of the first electronic literature showcase at the U.S. Library of Congress in 2013, Kathi Inman Berens and Susan Garfinkel selected sixty-nine books to contextualize the exhibit's twenty-seven featured works of electronic literature. This wide-ranging curatorial reach culled books that revealed various states of playfulness with the book's material form. While well-known examples such as *Choose Your Own Adventure* books and shaped poems ("concrete" poetry) emphasize unique aspects of playfulness and materiality, it was only through curating a set of books spanning hundreds of years and genres--and juxtaposing such books alongside the expressive interfaces of electronic literature--that a logic of playable books and electronic literature emerged. Not only did the books contextualize electronic literature, but the obverse happened as well: electronic literature's human/computer interface defamiliarized the book.

In the last years of the twentieth century, the printed book began to manifest attributes of digital art. In 1999 in the wake of "Booksday," an exhibition of unconventional books curated in Krakow by Polish writers and poets Zenon Fajfer, Katarzyna Bazarnik and Radoslaw Nowakowski, Fajfer suggested that books defying editorial conventions, whose linguistic content is inextricably bound with their material (printed) embodiments deliberately shaped by their authors, could be called "liberature." Initially juxtaposed with artists' books, liberature stresses the literary attributes of book-bound works, exploiting their semantically charged materiality. Although it began as a theory describing the codex form, liberature has strong links with electronic literature, through, for example, Fajfer's playable poems that are hybrid works combining the printed and digital interfaces, and Nowakowski's hypertextual, multimedia online narratives.

Also in 1999, N. Katherine Hayles published *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics*, a book that built the case for an interactive dynamic between seemingly disembodied information and the material substrates that convey them. These twin moves toward embodied experiences of reading suggest a parallel track between Polish and North American conceptions of reading abetted by experimentation in digital and networked environments. This is in tune with Jerome McGann's observation that "[t]he 'composition' of poetry is not completed--indeed, it has scarcely begun--when the writer scripts words on the page; and even at this initial moment of the imagination's work the scene is a social one. What kind of instrument is the writer using, what kind of paper?," which could be expanded to reflections about platforms, softwares, and mobile devices. These questions are central to understanding and interpretation of any work of verbal art, he continues, as they "are entangled with every textual network of meaning" (1993: 112). Each of the panelists will bring a specific focus to the history and future of playable books in theory and practice.

Susan Garfinkel will discuss examples from book history that display the mutability of the book as a version of Bakhtin's dialogic "world in the making" (1981). Laurence Sterne's *The Life and Opinions of Tristram Shandy, Gentleman*, for example, as early as 1759 made use of visual pastiche and typographic play along with its unusual plot, digressive presentation style, and literary borrowings. In 1804 and again in 1820, Thomas Jefferson famously cut apart and pasted up the Christian bible (Edwards, 2012). Working at the intersections of literary studies, media studies, and the digital humanities, Lisa Gitelman (2008) and Kirschenbaum (2008b), among others, have prompted media scholars to revisit the history of book materiality. Lori Emerson suggests that "by revisiting older media, we can make our current media visible once again" (2014: 130), while Kirschenbaum and Sarah Werner describe the digital within book history as "a frankly messy complex of extensions and extrusions of prior media and technologies" (2014: 408). Such awareness is enhanced by new scholarly attention to materiality, to the vibrancy of matter and the role of interpretation as mediation in the physical world (Trettien, 2013; Bennett, 2010; Appadurai, 2015). Looking back across decades and expressive genres for the precursors of electronic literature, we soon recognize an awareness of the dialogic mutability of the book itself, of its playable affordances that significantly predate the computer.

Katarzyna Bazarnik and Zenon Fajfer will focus on liberatic works and theory by exploring and expanding architecture of the codex, for example, in Oulipian *Cent Mille Milliards de Poèmes* by Raymond Queneau (1961), and their own , triple dos-à-dos of *Oka-leczenie* (2000, 2009), the book instrumental in defining the concept of "literature in the form of the book" in the central European context. However, liberature has also responded to digital technologies in various, subversive ways. In Fajfer's *ten letters* (2010), the poetic volume combining print and digital animation, it is paradoxically the material book that invites the reader to engage with the poems' visuality, to handle and manipulate the pages in a way reminiscent of Mallarméan 'espacement of reading,' ' *un espacement de la lecture*' (1998: 253). Its digital part, "Primum Mobile," accessible through the CD interface, is a piece blocked from the readers' intervention--they are only allowed to contemplate texts that infold and unfold in front of their eyes as an animated movie. In the spirit of playfulness its final section, "Spogladajac przez ozonowa dziure" ("Detect Ozone Whole Nearby") announces its further remediation as a "poem-in-the-bottle," with its text printed on a transparent plastic sheet placed in a glass container. Thereby, Fajfer's bottle-book seems to look back to the beginnings of book history, by taking on the form of the scroll. Lastly, his *Powieki* (Eyelids) (2013) is another bimedial work that seemingly, it returns to the traditional codex, yet textually is a densely linked hypertext that can be accessed via the printed or the electronic interfaces, offering readers radically different experiences of exploring the same cycle of poems.

Kathi Inman Berens will investigate the literary/ludic continuum in theories of ergodicity by Aarseth (1997), Ryan (2006), Laurel (2013) and Ensslin (2013) and examine how touch in electronic literature and playable books prompts new dimensions of liberatic engagement. Saleen and Zimmerman's *Rules of Play: Game Design Fundamentals* (2003) and the team of Robin Hunicke, Marc LeBlanc, Robert Zubek's MDA approach [Mechanics, Dynamics and Aesthetics] developed theoretical frameworks and critical language to unify conditions of play from board games to sports to computer and video games. Literary games don't figure in these discussions because they were developed before serious games initiated the hybrid literary game form. Jesper Juul's work on failure in games (2013) adds new analytic dimension to discussion of narrative, catharsis, pain, and reward systems pioneered in the work of Janet Murray (1999; 2012). Can one "lose" when playing a work of literature? Serge Bouchardon's "Loss of Grasp" (2010), winner of the 2011 New Media Writing Prize, has elicited scholarly studies among Polish literary critics applying liberatic theory and English-speaking literary critics examining literary games. A comparison of these methods could frame cross-cultural ways to analyze new works such as Inkle Studio's *80 Days* (2014), a tablet game based on Jules Verne's *Around the World in Eighty Days*, and Steve Tomasula's *TOC* (2009, 2013). *80 Days* was *Time Magazine*'s 2014 Game of the Year and also, tellingly, *The Guardian*'s 2014 Novel of the Year. This generic slippage between book and game is likely to become more common as book publishers venture into gameful environments, such as Doubleday's *Bats of the Republic* (2015), a playable novel featuring haptic elements. Berens analyzes the material book's ludic dialog with tablet-based digital stories.

Following a presentation of these varied approaches to the questions of playability and interface in books past and present, we plan a dialogue between ourselves and the audience. We conceive our conversation around the concept of playable books as a fertile starting point for thinking toward an expansively multiple understanding of agency and affordance in reading-playing, centered in acts of encounter for both the author-designer-creator and the reader-user-player. The ludic, performative materiality of all books, analog or digital, opens up inquiry across a broad range of hybrid instantiations: from children's digital literature including the Sony Wonderbook works, *The Sailor's Dream* (Simogo, 2015), and the *Mrs. Wobbles* series (Marino Family, 2013-present); to hybrid mass market haptic books like *Tree of Codes* (Foer, 2010), *Nox* (Carson, 2010) and *S.* (Abrams and Dorst, 2013); to app-based artists' books such as *Between Page and Screen* (Borsuk and Bouse, 2012), *Abra* (Borsuk, Durbin and Hatcher, 2014) and *Pry* (Tender Claws, 2013).

Bibliography

Primary sources

Abrams, J. and Dorst, D. (2013). *S.* London: Mulholland Books.

Bazarnik, K. and Fajfer, Z. (2003). *(O)patrzenie*. Liberatura series, vol. 1. Kraków: Krakowska Alternatywa (renamed Korporacja Ha!art).

Borsuk, A. and Bouse, B. (2012). *Between Page and Screen*. Los Angeles; New York: Siglio.

Borsuk, A., Durbin, K. and Hatcher, I. (2014). *Abra.* http://www.a-b-r-a.com/ .

Bouchardon, S. (2010). *Loss of Grasp*. http://lossofgrasp.com/ (accessed 5 March 2016).

Carson, A. (2010). *Nox*. New York: New Directions.

Danielewski, M. Z. (2000). *Mark Z. Danielewski's House of Leaves*. New York: Pantheon Books.

Dodson, Z. T. (2015). *Bats of the Republic*. New York: Doubleday/Knopf.

Foer, J. (2010). *Tree of Codes*. London: Visual Editions.

Fajfer, Z. and Bazarnik, K. (2000, 2009). *Oka-leczenie*. Liberatura series, vol. 8. Kraków: Korporacja Ha!art.

Fajfer, Z. (2013). *Powieki* (Eyelids). [Print, CD and online] Szczecin: Forma. http://techsty.art.pl/powieki/ (Accessed 5 March 2016).

Fajfer, Z. (2004, 2009). *Spoglądając przez ozonową dziurę* [Detect Ozone Whole Nearby]. 2d ed. Liberatura series, vol. 2. Kraków: Korporacja Ha!art.

Fajfer, Z. (2010). *Ten letters*. Ed. and trans. by K. Bazarnik. Liberatura series, vol. 11. [Print and DVD.] Kraków: Korporacja Ha!art.

Fisher, C. (2013). *200 Castles*. https://projeqt.com/caitlin/200-castles.

Fisher, C. (2012). *Circle*. https://vimeo.com/64504258.

Inkle Studios. (2014). *80 Days*.http://www.inklestudios.com/80days/.

Lewis, J. E. and Nadeau, B. (2008-2013). *P.o.E.M.M. = Poetry for Excitable [Mobile] Media*. Writing Complex. http://www.poemm.net/.

Marino Family. (2013-present). *Mrs. Wobbles & the Tangerine House*. http://markcmarino.com/mrsw/ .

Nowakowski, R.*Liberatorum. A Book Laboratory.*http://www.liberatorium.com/ (accessed 5 March 2016).

Queneau, R. (1961). *Cent Mille Milliards de Poèmes*. Paris: Gallimard.

Simogo. (2015). *The Sailor's Dream*. http://simogo.com/work/the-sailors-dream/ .

Sterne, L. (1779). *The Life and Opinions of Tristram Shandy, Gentleman: In Three Volumes*. Dublin: printed for J. Potts, J. Williams, W. Colles, T. Walker, C. Jenkin, and L. White.

Tender Claws. (2013). *Pry.* http://prynovella.com/ .

Tomasula, S. (2009, 2013). *TOC: A New Media Novel*. University of Alabama Press.

Secondary sources:

Aarseth, E. J. (1997). *Cybertext: Perspectives on Ergodic Literature*. Baltimore: Johns Hopkins UP.

Appadurai, A. (2015). Mediants, Materiality, Normativity. *Public Culture*. doi:10.1215/08992363-2841832 (accessed 29 August 2015). **27**(2 76): 221–37.

Bakhtin, M. M. (1981). *The Dialogic Imagination: Four Essays*. (Ed.) Holquist, M. (Trans.) Emerson, C. and M. Holquist. Austin: University of Texas Press.

Bennett, J. (2010). *Vibrant Matter: A Political Ecology of Things*. Durham: Duke University Press.

Berens, K. I. (2015). Touch and Decay: Porting Tomasula's *TOC* to iOS. In: *The Art and Science of Steve Tomasula's New Media Fiction*, ed. Banash. New York: Bloomsbury, pp. 167-82.

Cramer, F. (2016). Post-Digital Literary Studies. *MATLIT: Revista Do Programa de Doutoramento Em Materialidades Da Literatura*. doi:10.14195/2182-8830.**4**(1): 11–27.

Drucker, J. (2013). Performative Materiality and Theoretical Approaches to Interface. *DHQ*. http://digitalhumanities.org/dhq/vol/7/1/000143/000143.html (accessed 31 October 2015). **7**(1).

Edwards, O. (2012). How Thomas Jefferson Created His Own Bible. *Smithsonian Magazine*, January. http://www.smithsonianmag.com/arts-culture/how-thomas-jefferson-created-his-own-bible-5659505/ (accessed 31 October 2015).

Emerson, L. (2014). *Reading Writing Interfaces: From the Digital to the Bookbound*. Minneapolis: University of Minnesota Press.

Ensslin, A. (2014). *Literary Gaming*. Cambridge: The MIT Press.

Fajfer, Z. (1999). Liberatura. Aneks do słownika terminów literackich. *Dekada Literacka* 5-6 (30 June), pp. 8-9.

Fajfer, Z. (2010). *Liberature or Total Literature. Collected Essays 1999-2009*. (Ed. and trans.) Bazarnik, K. Liberatura series, vol. 12. 1st ed. [pdf] Kraków: Korporacja Ha!art. http://www.ha.art.pl/e-booki/Zenon_Fajfer_-_Liberature_or_Total_Literature_ENG.pdf (accessed 31 October 2015).

Garfinkel, S. (2013). From Books to Bits: Library of Congress Electronic Literature Showcase Highlights Emerging Literary

Forms. *The Signal: Digital Preservation.*http://blogs.loc.gov/digitalpreservation/2013/04/from-books-to-bits-library-of-congress-electronic-literature-showcase-highlights-emerging-literary-forms/ (accessed 20 October 2015).

Gitelman, L. (2008). *Always Already New: Media, History, and the Data of Culture*. Cambridge: The MIT Press.

Hayles, N. K. (1999). *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press.

Hayles, N. K. (2002). *Writing Machines*. Cambridge: The MIT Press.

Hayles, N. K. (2008). *Electronic Literature: New Horizons for the Literary*. Notre Dame, Ind: University of Notre Dame.

Hunicke, R., LeBlanc, M. and Zubek, R. "MDA: A Formal Approach to Game Design and Game Research." http://www.cs.northwestern.edu/~hunicke/MDA.pdf (accessed 5 March 2016).

Juul, J. (2013). *The Art of Failure*. Cambridge: The MIT Press.

Kalaga, W. (2010a). Liberature: Word, Icon, Space. In: Z. Fajfer, *Liberature or Total Literature. Collected Essays 1999-2009*. (Ed. and trans.) Bazarnik, K. 1st ed. Kraków: Korporacja Ha!art, pp. 9-19.

Kalaga, W. (2010b). Tekst hybrydyczny. Polifonie i aporie doświadczenia wizualnego. In: (Eds.) Bolecki, W and Dziadek, A., *Kulturowe wizualizacje doświadczenia*. 1st ed. Warsaw: IBL and Fundacja "Centrum Międzynarodowych Badań Polonistycznych," pp. 74-104.

Kirschenbaum, M. G. (2008a). Bookscapes: Modeling Books In Electronic Space. *Human-Computer Interaction Lab 25th Annual Symposium*. https://mkirschenbaum.files.wordpress.com/2013/01/bookscapes.pdf .

Kirschenbaum, M. G. (2008b). *Mechanisms: New Media and the Forensic Imagination*. Cambridge, Mass: The MIT Press.

Kirschenbaum, M. and Werner, S. (2014). Digital Scholarship and Digital Studies: The State of the Discipline. *Book History*, **17**(1): 406–58.

Laurel, B. (2013). *Computers as Theatre*, 2d ed. New York: Addison-Wesley Professional.

Mallarmé, S. (1998). *Poésies et autres textes*. Paris: Le Livre de Poche.

Malopolska Institute of Culture. (2009). *(O)patrzenie by K. Bazarnik and Z. Fajfer (2003, Liberatura). Commentary by B. Zalewski.*https://www.youtube.com/watch?v=-Z_Etuv_cl4 (accessed 31 October 2015).

McGann, J. (1991). *Textual Condition*. Princeton: Princeton University Press.

McGann, J. (1993). *The Black Riders. The Visible Language of Modernism* . Princeton: Princeton University Press.

Murray, J. (1999). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Cambridge: The MIT Press.

Murray, J. (2012). *Inventing the Medium: Principles of Interaction Design As a Cultural Practice*. Cambridge: The MIT Press.

Rowe, S. (2012). *Fantasies of Contact: Erica Baum, Susan Howe, and the Poetics of Paper*. http://www.full-stop.net/2012/06/27/features/essays/sam/fantasies-of-contact-erica-baum-susan-howe-and-the-poetics-of-paper/ (accessed 6 November 2015).

Ryan, M.-L. (2006). *Avatars of Story*. Minneapolis: University of Minnesota Press.

Saleen, K. and Zimmerman, E. (2003). *Rules of Play: Game Design Fundamentals*. Cambridge, Mass: The MIT Press.

Trettien, W. A. (2013). A Deep History of Electronic Textuality: The Case of English Reprints John Milton Areopagitica. *Digital Humanities Quarterly*, **7**(1). http://digitalhumanities.org/dhq/vol/7/1/000150/000150.html (accessed 1 November 2015).

Wardrip-Fruin, N. and Harrigan, P. (eds). (2004). *First Person: New Media as Story, Performance, and Game*. Cambridge, Mass: The MIT Press.

Institutionalizing and implementing the Digital Yiddish Theatre Project

Joel Berkowitz
berkowit@uwm.edu
Digital Yiddish Theatre Project, United States of America

Debra Caplan
Debra.Caplan@baruch.cuny.edu
Digital Yiddish Theatre Project, United States of America

Agnieszka Legutko
a.legutko@columbia.edu
Digital Yiddish Theatre Project, United States of America

Aaron Rubinstein
arubinst@library.umass.edu
Digital Yiddish Theatre Project, United States of America

Administrative Challenges of a Global Collaborative Research Group

Joel Berkowitz

DYTP co-founder Joel Berkowitz will address administrative challenges facing a collaborative research group. These include the initial selection of group members, the division of labor within the group, articulating the group's mission and objectives, keeping activities moving forward in a timely fashion, managing effective communication among group members spread across a number of cities on two continents, working with partner individual organizations, and funding the consortium's work.

Data and the 'Obscurity' of Yiddish Theatre

Debra Caplan

DYTP co-founder Debra Caplan will describe the group's progress in creating a data model and database for a digital critical edition of a Yiddish theatre encyclopedia. Yiddish theatre data poses particular challenges, and our data model and database will need to accommodate relationships among individual actors, directors, and playwrights; hundreds of itinerant Yiddish theatre companies; thousands of geographical locations, often within rapidly changing national borders; a multilingual repertoire that often featured dozens of variants of a single play; literary trends in Yiddish drama; ticket prices and sale figures; cross-continental reviews; and other data related to the near-constant travels of Yiddish theatre artists across the globe. Caplan will discuss the potential of this data to document Yiddish theatre's broad impact on theatre history.

Capturing the Ephemeral: The Digitization of Yiddish Theatre

Agnieszka Legutko

Agnieszka Legutko will set the Digital Yiddish Theatre Project in a broader context of other Yiddish theatre-related digital attempts, often short-lived or only partially executed online projects created in the last decade or so. She will also discuss the unique potential of the Digital Yiddish Theatre Project resulting from its focus on specific content, collaboration of a diverse team of experts, and strategies for attaining social media presence.

Creating a Yiddish OCR

Aaron Rubinstein

Data mining, the computer analysis of large corpora, is an essential tool in contemporary textual analysis. For the humanities, Optical Character Recognition (OCR) has made this kind of analysis possible, rendering the text from printed, digitized works into a machine readable form. Though far from perfect, OCR has a high success rate recognizing roman characters but for researchers in the field of Yiddish studies, OCR engines that can recognize characters from the Jewish alphabet are still in the early stages of development. This presentation will discuss the current state of Yiddish alphabet OCR and a project that attempts to make use of existing tools and a variety of other strategies to provide machine readable access to a seminal Yiddish reference work.

Linked Ancient World Data: Relating the Past

Gabriel Bodard
gabriel.bodard@sas.ac.uk
University of London, United Kingdom

Tom Gheldof
tom.gheldof@arts.kuleuven.be
KU Leuven

K. Faith Lawrence
faith.lawrence@kcl.ac.uk
King's College London

Simona Stoyanova
simona.stoyanova@informatik.uni-leipzig.de
University of Leipzig

Charlotte Tupman
c.tupman2@exeter.ac.uk
University of Exeter

Leif Isaksen
l.isaksen@lancaster.ac.uk
University of Lancaster

Rainer Simon
rainer.Simon@ait.ac.at
Austrian Academy of Sciences

Elton Barker
elton.barker@open.ac.uk
Open University

Pau de Soto Cañamares
paudsoto@gmail.com
Institute of Catalan Studies

Hugh Cayless
hugh.cayless@duke.edu
Duke University

Bringing together researchers from a wide variety of disciplinary backgrounds, this panel discusses the newly emerging ecosystem of Linked Ancient World Data projects and resources. One of the fastest growing areas of Digital Humanities, Linked Open Data (LOD) has the potential to transform traditional scholarship through its ability to promote the discovery of, and connections between, online documents of a highly varied nature (texts, maps, databases, images, etc.). Yet many barriers that are limiting its uptake and application, both technical and human, need to be addressed before this potential can be realised. This panel explores the issues relating to LOD, the semantic web and RDF technologies by focusing on three case studies drawn from the ancient world of literature, history and archaeology: the SNAP, Pelagios, and Integrating Digital Epigraphies (IDEs) projects.

Each of these three projects takes a different focus for their linking strategies: SNAP aims to connect documents through the people mentioned within them (prosopographies and onomastica); Pelagios through places (maps and gazetteers); and IDEs tackles different kinds of written material that survive from the ancient world (inscriptions and papyri). The projects are all united, however, in a concern for the use of, and access to, massive and diverse datasets that cannot be curated, aggregated or even archived in a single location. One major challenge to be addressed is the inherited scholarly infrastructure that tends to shoehorn multiple projects into a single institution's repository and data model. These projects and their participants are also concerned with issues far beyond their primary subject area: the interoperability of bibliographical references, citations of ancient sources, encoding of date and time, events and actors, material objects and their curatorial history all contribute to the study and understanding of the ancient world (and *mutatis mutandis* of any other). All also recognise that there is no firm demarcation between the cultures of the Mediterranean in the classical period, nor between the worlds and cultures bordering them in time and space. Data from the Bronze Age and Mediaeval periods can be read profitably in and against this classical focus; our sources do not exist in a vacuum that can be entirely insulated from the ancient Near and Far East, sub-Saharan Africa or pre-Columbian America, for example. The three papers in this panel will all discuss how they are deploying the formats and technologies of Linked Open Data to address the massive and multidisciplinary interoperability that these historical challenges require.

Networking Ancient Person-data: community building and user studies around the SNAP:DRGN project

Gabriel Bodard
Tom Gheldof
Faith LawrenceSimona Stoyanova
Charlotte Tupman

The Standards for Networking Ancient Prosopographies (SNAP: http://snapdrgn.net) project is using linked open data (LOD) to build a virtual authority list for ancient people through aggregation of common information from collaborating projects. A unified authority of ancient persons will serve as a convenient and powerful single resource for prosopographers, text editors and scholars to use for disambiguating person references by means of annotations that record the specific URI of a person identified by the SNAP graph.

The objective is neither to create a new universal dataset of historical persons, nor to ingest or supplant the many valuable prosopographical resources, both analogue and digital, created over the past many years. Rather, through the creation of a single entry point—and related identifier—coupled with a small subset of common fields made available both to human researchers and for automated processing, SNAP aims to facilitate interoperability and interchange, exploitation and discovery through common metadata, and the recording of both known and newly discovered relationships between person records. Users will be enabled and encouraged to (a) annotate their data with SNAP URIs to disambiguate person references, and (b) add structured commentary to the SNAP graph in the form of scholarly assertions, bibliography and apparatus.

This paper will outline our efforts to engage both the scholarly community and the wider public in the development of the SNAP model, and discuss the importance of user analysis and feedback into the design and functionality of the user interface and research tools. It is essential both for the utility of the project, and to encourage scholarly uptake in the person data and use of the virtual authority records, that we base our development on consultation with communities of our anticipated user groups, both scholarly and more widely.

In the first phase of the project, SNAP began to address the core issue of linking together large datasets containing information about persons, names and person-like entities (families, associations, deities, anthropomorphic animals) managed in heterogeneous systems and formats. Ambiguous co-referencing is a ubiquitous issue within the linked data world; how does a researcher or analyst determine whether two records refer to the same person or are related in some other way? Even more trickily, what other related information referring to one record can be said equally to apply to both people? The SNAP dataset attempts to address this issue by retaining scholarly metadata around the assertion of co-references and relationships, so that ambiguity, disagreement and academic justification can be recorded alongside all statements that can potentially lead to inference of new relationships.

SNAP models a simple structure using Web and LOD technologies to represent relationships between databases and to link from references in primary texts to authoritative lists of persons and names. The core of its source material was built around three large historical prosopographies and onomastica (databases of persons and names) from the ancient world: the *Lexicon of Greek Personal Names*, an Oxford-based corpus of some 300,000 persons mentioned in ancient Greek texts (http://www.lgpn.ox.ac.uk/); *Trismegistos*, a Leuven-run database of over half a million names and persons from Egyptian documents (http://www.trismegistos.org/); and *Prosopographia Imperii Romani*, a series of printed books listing senators and other elites from the first three centuries of the Roman Empire (http://pir.bbaw.de/). With several other more specialist databases of ancient and Mediaeval persons, museum and library catalogues, and digital editions contributing data or in the process of converting their data to the RDF format SNAP requires, the virtual authority will soon record well over a million person URIs. Due to the focus on historical datasets, we are able to address wider issues of dealing with person data without the ethical and privacy concerns raised by that gained from modern social networks. While still massive in scale, the amount of data under discussion is tractable, allowing for more academic coherence and review within the data, which, diverse as it is, is produced by a discipline with well-established working practices.

During this first phase, SNAP held a number of meetings and presentations to introduce the principles of and the preliminary work done by the project in its pilot period, and to hear from potential project partners about their datasets, practices and reactions to our proposals. These discussions led to a greater understanding of the nature of the prosopographical materials as well as helping to identify future partners. In addition, they provided the opportunity to advise participants how to present the relevant subset of their data for SNAP import in order to allow further datasets to be ingested, by demonstrating the *SNAP Cookbook* (http://snapdrgn.net/cookbook/) which sets out details of several scenarios for the encoding, publication and linking of ancient person data in RDF, and connecting them to the SNAP graph.

The second phase of the SNAP project focuses further on the ingest, creation and linking of a much wider range of person data, using a range of methodologies including named entity recognition (NER), both hand and machine-assisted curation of person references from large corpora including inscriptions and mythological sources, and the ingest of data from existing projects via prosopographical tool kits such as the Berlin-Brandenberg Academy's *Personendaten Repositorium* (http://pdr.bbaw.de/) and the Berkeley Prosopography Services (http://berkeleyprosopography.org/).

At the same time, we aim to enable a wider interchange of data and discovery of related materials as part of the larger Linked Ancient World Data community. In order to engage scholars and other interested parties in both creating and linking to SNAP identifiers, and to support and encourage the use of research tools and interfaces, we will hold a series of user analysis and engagement workshops, focusing on scholarly feedback on the SNAP web interface, API and widgets, the use of disambiguating annotation, and creation of and engagement with structured commentary. These workshops will help us assess and better understand the expectations and needs of our user communities with regards to both infrastructure and support, and sustained engagement with the project. User analysis is vital for developing an understanding of how users interact with the data and the current tools that are available for working with prosopographical data. We need to understand the goals and workflows of our user groups in order to meet their scholarly needs, and to create effective methods for attracting and maintaining engagement with the next phase of the project.

The issues being investigated by SNAP have implications beyond the bounds of this particular project: many of the issues such as variant name spellings, persons with changing or ambiguous names, uncertain identities and relationships, and tracking assertions about persons and the cascading inferences resulting from scholarly or editorial decisions, are precisely the same questions that concern both professional and amateur groups working on person-identification, including local historians, family historians, genealogists and graveyard conservationists. This has two major implications: first, as the work being done by SNAP is likely to reach far beyond its immediate subject area, this must be reflected in the way the project is conducted and disseminated to a variety of groups; and second, we must ensure that our user engagement workshops therefore include not only scholars but members of the public whose interests overlap with those of the SNAP project in these and other ways. This paper will seek audience discussion and advice about how best to ensure that the second phase of the project can meet such potentially diverse user needs.

Early Geographic Documents and the Pelagios Commons

Leif Isaksen
Rainer Simon
Elton Barker
Pau de Soto Cañamares

Pelagios is an international initiative concerned with the development of LOD methods, tools and services so as to better interconnect the vast and ever-growing range of historical resources online. Specifically, it uses the Open Annotation RDF ontology (http://www.openannotation.org/spec/core/) to associate place references within those resources to online gazetteers that offer URI-based identifiers for such places. The resulting graph is then exploited in a variety of ways to facilitate research, teaching and public engagement. The Pelagios 3 project expanded the scope of Pelagios dramatically from its original focus on classical antiquity, to encompass the early geographic documents of the pre-modern era, including early Christian, Islamic and Chinese traditions. It addressed three critical challenges for stimulating activity in these areas:

First, we developed user-friendly Web-based and Open Source software tools for the production and exploration of Pelagios LOD. Recogito (http://pelagios.org/recogito/) is a Web-based tool for the semi-automatic annotation of place references. It features several work areas, dedicated to different stages of the geo-annotation workflow: (i) a text annotation area to identify place names in digital texts or tabular documents (optionally aided by automatic Named Entity Recognition); (ii) an image annotation area to mark up and transcribe place names on high-resolution map or manuscript scans; (iii) a geo-resolution area, where identified (and transcribed) place names are mapped to a gazetteer, supported by an automated suggestion system. Recogito also provides basic features for managing documents and their metadata, as well as for viewing annotation results, usage statistics and bulk-downloading annotation data. Peripleo (http://pelagios.org/peripleo/map) is a spatio-temporal search engine for exploring the annotation data produced through Pelagios 3, as well as by the Pelagios community at large. Its user interface resembles that of Google Maps, and allows for free browsing as well as keyword & fulltext search, while offering additional filtering options based on time, data source and object type.

Second we carried out much annotation both in house and by independent contributors, so as to provide a 'critical mass' of annotated text and map documents that would attract contributions from other data curators. Over the course of the project 90 registered editors identified approximately 130,000 place references in 317 early geographic documents in 8 languages. About half of these were manually inspected for association with a gazetteer. Around 60 institutional or personal partners have contributed to Pelagios to date with a similar number expressing interest in doing so. We believe this offers substantial evidence that LOD approaches do not of necessity impose high barriers to entry, and on-ramps to semantic technologies can be offered at a varying levels of complexity.

Third, we developed a mechanism for enabling different gazetteers (each serving their particular community) to be interoperable, allowing for interlinking between data from divergent traditions. This has been achieved through the development of the Pelagios Gazetteer Interconnection Format which provides baseline requirements and optional additions for gazetteers to interoperate (https://github.com/pelagios/pelagios-cookbook/wiki/Pelagios-Gazetteer-Interconnection-Format). While such decentralized models for key infrastructure are both evolving and not without their challenges and risks, they offer significant potential for resolving conventional problems with enforcing universal standards across multiple domains and communities of practice.

Consequently, Pelagios has generated sustained and lively community interest, and has offered a pioneering model for other LOD initiatives which are semantically annotating different reference types from people to time periods, including PeriodO (http://perio.do/), SNAP (https://snapdrgn.net/), PastPlace (http://www.pastplace.org/) and al-Thurayya (http://maximromanov.github.io/projects/althurayya_02/). The success of the Pelagios approach has also attracted funding for academic research into early geographic documents through the Pelagios 4 project which is working with specialists in historical geography to identify both the advantages and limitations of semantic annotation for comparative studies and visual and statistical analyses. Topics span from the significance of hazard depictions on medieval portolan charts to the use of reliability of textual sources as proxies for the missing sections of the only extant Roman world map. In addition to this academic research, the SEA CHANGE project trialled crowdsourcing workshops for the use of semantic annotation in Higher Education in collaboration with i3 Mainz and the University of Heidelberg (http://pelagios-project.blogspot.co.uk/2014/11/bringing-about-sea-change.html).

In parallel with these developments a community of practitioners has emerged with interests in a range of related activities: the annotation of curated or third-party content; the production of specialist gazetteers; the integration of place annotations with those of people, periods and things; and the visualization and analysis of graph-based data, to name but a few. Since its early stages Pelagios has made concerted efforts to consult and support such stakeholders, but as it has grown new opportunities and challenges have emerged. In particular we have established that within a heritage context, one of LOD's key advantages is its ability to relate independently maintained projects without requiring a single centralized authority. But what are the social ramifications of such an approach? In a world in which funding criteria, academic legitimacy, intellectual property, and even conference presentations presume the authority of individuals and institutions, can LOD communities ever scale effectively? In order to do so individual stakeholders will need to shoulder responsibility for specific services within them. These may be the provision of content, real-time search aggregators, or dynamic real-time operations that offer visualization and analysis of heterogenous material. For some, reliability will be more important than complexity or innovation, while others will pioneer new strategies at the cost of longevity or broad usership. Negotiating the relationships between these stakeholders—technically, socially and legally—is perhaps the greatest task ahead for those seeking to establish Linked Open Data as a principal mechanism for drawing together, if not necessarily synthesizing, information about the past.

In addition to reviewing the outputs of Pelagios 3 and Pelagios 4, this paper will report on early developments within Pelagios Commons, a new phase of Pelagios which focuses explicitly on increasing its technical and social decentralization. This spans beyond its current pre-Modern and literary scope, in order to embrace later periods, differing scales of geography (from intra-urban to multi-regional) and the conceptual changes of dealing with arbitrary findspots and mythical, fictional and itinerant places. It will present our experiences in establishing Special Interest Groups, and the different challenges faced in devolving LOD architectures, as well as lessons learned from similar initiatives. In doing so we hope to foster discussion and critique from those planning or implementing related community-driven projects.

Acknowledgments: Pelagios 3 was funded by the Andrew W. Mellon Foundation. Pelagios 4 was funded by

the AHRC. SEA CHANGE received an Open Humanities Award from D2ME and the Open Knowledge Foundation. We thank all our financial supporters and content contributors for the collective production of this work.

Integrating Digital Epigraphies

Hugh Cayless

The Integrating Digital Epigraphies (IDEs) project aims to build on the lessons learned in the course of developing the Papyri.info project. The differences in the digital landscape between Greek Epigraphy and Papyrology are considerable, the main one being that, whereas many of the partner projects of Papyri.info were happy to permit that site to aggregate their data, IDEs will not be able to host partner data, but rather to collect citation and linking information with the goal of improving the links between the different Epigraphical sites. This paper will discuss the project, hosted athttp://ides.io and the part played in it by the Linking Ancient World Data Ontology, which may be found athttp://lawd.info.

What is a citation? It is a sequence of characters in a text that refers to something the reader may wish to consult. A citation is obviously a pointer of some sort, but what is it a pointer to? It depends: a citation like *Il 1.1* is a reference to book 1, line 1 of the *Iliad*, that is, it refers, notionally, to an ideal or composite Iliad, not to a particular expression. It is assumed that the first line of book 1 will be more or less the same in any edition. A citation like *IG I^3 40* on the other hand, is to a particular edition: number 40 in the third edition of the first volume of *Inscriptiones Graecae*. It points therefore to a particular part of a larger work. *S.C. de Bacch.*, to take a third example, is a citation of an actual inscription, not a publication of that inscription. Here, the citation refers to a physical object, and the reader is expected to be able to find a text of it if they wish to read the document.

Given all this, we must conclude that if a citation is a pointer, it is a very vague one. It may point to an abstract work, an actual (perhaps online) edition or part of it, or a real-world object with text written on it, but which the reader of the citation probably isn't expected to go and read in person. Further complicating the situation, the referring strings that comprise citations are subject to different formatting conventions: *IG I^3 40*, and *IG I[3].40* refer to the same edition, for example. Even more extreme variation is entirely possible, depending on the conventions used by the publications citing the source.

How do we model this situation then? Citations are strings that indicate resources or parts of resources, which may or may not be abstractions and may or may not have published editions, translations, etc.. The LAWD Ontology approaches this by modeling the Citation as its own RDF Class which may represent a written or conceptual work. Citations may have a value that is either a string, a URI, or both, and other properties may be attached to them as well.

In IDEs, Citations are the main component of the project. IDEs works by identifying epigraphic citations from a variety of sources, including the Packard Humanities Institute (PHI) site, the Diccionario Griego-Español's Claros project, and the Supplementum Epigraphicum Graecum (SEG). PHI contains the text of editions of Greek inscriptions, Claros collects pairs of citations where one source cites or updates another, and SEG publishes a kind of annotated bibliography of epigraphic publications. All of these manifest different citation practices, even though they deal with largely the same body of material. IDEs attempts to parse the citations from each project, match them up when that is possible, and present an interface and APIs that permit projects with epigraphic citations to retrieve related material easily. For example, if PHI wishes to find and display citations related to the inscription they assign ID number 40 to (*IG I^3 40*), they can query the URI http://ides.io/browse/ides:phi:40 which resolves to http://ides.io/browse/ides:t00003n (the IDEs ID or IDEst of the inscription itself), and from there they can retrieve machine readable data in JSON, RDF, or JSON-LD formats that could be incorporated into their own page at http://epigraphy.packhum.org/text/40. It would be possible, for example, to link to related articles in SEG, to display additional bibliography, to link to the corresponding place entry in Pleiades, and when we have incorporated data from JSTOR, to link to articles that mention *IG I^3 40*.

IDEs itself is a property graph database which uses RDF semantics without relying on an underlying RDF database implementation. This approach is intended to allow it to permit commenting on and editing relationships between the entities it tracks. RDF by itself has trouble attaching extra information to triples, requiring reification or the use of named graphs to do so. We hope that by imposing RDF semantics on a property graph structure, we can have the ability to attach additional metadata and commentary to relations without sacrificing speed or expressive capability.

Creating Feminist Infrastructure in the Digital Humanities

Susan Brown
sbrown@uoguelph.ca
School of English and Theatre Studies, University of Guelph, Canada; English and Film Studies; Humanities Computing, University of Alberta, Canada

Tanya Clement
tclement@ischool.utexas.edu
University of Texas at Austin, USA

Laura Mandell
mandell@tamu.edu
Texas A and M University, USA

Deb Verhoeven
deb.verhoeven@deakin.edu.au
Deakin University, Australia

Jacque Wernimont
jacqueline.wernimont@asu.edu
Arizona State University

> Today, it is imperative that we develop an ideological infrastructure that both supports and facilitates feminist interventions within connective, networked elements of the contemporary world. [...] We want to cultivate the exercise of positive freedom – freedom-to rather than simply freedom-from – and urge feminists to equip themselves with the skills to redeploy existing technologies and invent novel cognitive and material tools in the service of common ends.
>
> (Laboria Cubonix, Xenofeminism: A Politics for Alienation)

Introduction

This panel considers how gender and digital infrastructures shape each other. Infrastructure can be described as that which creates the conditions of possibility for certain kinds of activities. In the context of digital humanities, infrastructure can refer to physical infrastructure such as servers, software infrastructure in the form of code or a software stack, organizational infrastructure such as a scholarly society, institutional infrastructure such as a DH centre, or methodological infrastructure such as a standard.

As is often observed, infrastructure tends to be transparent or invisible until broken (see Bowker and Starr), rather in the manner that ideological structures can blind us to systemic discrimination and gender bias. Rather than view infrastructure as a transparent mediation, feminist thinking invites broader questions around how we might address the social or relational aspects of infrastructure:

- How can digital infrastructure, as technologies of connection, support complex, non-binary understanding?
- How does a systemic approach to information infrastructure offer opportunities for a more systemic political critique within the digital humanities?
- How can we address the ways in which standards (such as the TEI encoding of sex [Terras]) and procedures ('organizing logic' [Posner]) embed values?
- What would an explicitly 'ideological infrastructure' look like? Is it desirable?
- To what extent can existing infrastructure be adapted for feminist ends and/or do we need to create new forms and instances of infrastructure to redress inequity?

The panel will be a hybrid of the panel- and multiple-paper session. Within each of its three sectioned themes, the panelists specified for a particular subtopic will have the responsibility to open up discussion with short statements of up to 7 minutes, followed by pithy responses from other panel members (who will have read the statements in advance) of up to 5 minutes in total, followed by an invitation to the audience to continue the discussion jointly with us. What is proposed, therefore, is not a series of 5 papers, but a set of interweaving engagements with each other and our audience that will make clear the connections between the component topics.

Training and pedagogical traditions

DH Training (Tanya Clement) Project-based research is often heralded as the site of work that defines DH. Indeed, in Willard McCarty's DH paradigm, the DH practitioner alone possesses 'an outsider's objectivity' and the 'performative ability to move in and out of disciplines, back and forth between duck and rabbit... while carrying its own intellectual load on its back' (136). Except, a feminist perspective based on situated knowledges (Haraway, 1998) teaches us that what liberates us from obstacles often blinds us to the opportunities that difficulties illuminate.

This rhetoric of 'mastery' over technology threatens an advancement of knowledge production from other perspectives, adopting what Haraway would call 'the standpoint of the Man, the One God, whose Eye produces, appropriates and orders all difference' (Haraway, 1988: 587). Further, training in project-related work arguably produces skills- rather than knowledge-based programs, and digital humanists professionals rather than researchers (Clement, 2012; Kraus, 2013; Mahony and Pierazzo, 2012). The debate about whether designing and building tools provides indispensable knowledge about information infrastructures is inflected by the concern that the ability to bridge the 'building' gap) largely reflects social privilege.

Clement will posit a contradiction between the presumed goals behind DH work, and the everyday practices of DH scholars and major DH education and training programs. In short, while scholars claim a keen desire to frame infrastructure development in the context of theories such as cultural criticism, feminist inquiry, and post-colonial critique, her investigation indicates that many training programs are not framed in these ways.

Community-based Training and Research Networks (Jacque Wernimont): FemTechNet is an activated network of scholars, artists, and students working on, with, and at the borders of technology, science, and feminism in many fields including Science and Technology Studies, Digital Humanities, Media and Visual Studies, Art, Gender, Queer, and Ethnic Studies. This networked structure works asynchronously and across local and international interventions. A major feature of our work is the series of Distributed Open Collaborative Courses (DOCCs) that recognize the complexities of the learning situation by designing platforms both locally and collectively. They represent new models from which many stakeholders will learn. The DOCCs are alternatives to the 'reform' efforts represented by the Massive Open Online Courses (MOOCs), which actually do little to change the status quo and can even be counterproductive when promoters oversimplify questions of access, underestimate investments of labor in instructional technology, deny the importance of infrastructure and its human and discursive aspects, and reinforce ideologies about technology being values-neutral.

In addition to curricular work, FemTechNet members collaborate on the design and creation of feminist technological innovations and in so doing extend our networked training into both traditional curriculum, apprenticeship, and non-hierarchical collaboration. Consequently, FemTechNet is a model of research and training infrastructure that foregrounds transparent and locally responsive structures, affords the ability to move up and back as time, expertise, and desire allow, and places feminist principles of equity, justice, and engagement at the fore. It is also a case study in understanding the labor, care, and resources necessary to maintain international cooperation and collaboration. We believe that our infrastructure is as expressive of our intellectual and ethical commitments as any content that we might produce, making the challenges of distributed online work around digital technologies not simply innovative, but also absolutely necessary.

Examples of feminist technical infrastructure

To give a sense of the array and complexity of infrastructural concerns, we will frame three quite different technical infrastructure initiatives in relation to feminism.

HuNI (Deb Verhoeven): HuNI, Humanities Networked Infrastructure, is a service that aggregates data from a broad range of humanities disciplines and institutional sources and makes them available to researchers and any members of the public with access to a browser (Verhoeven and Burrows, 2015). HuNI provides users the opportunity to build collections and propose a graph of relationships between records which in turn contribute to a HuNI network graph. As researchers themselves create the links between data, they also produce a kind of 'vernacular ontology' which, rather than providing an 'authoritative' model of the data, instead allows for diversity, complexity, interpretation and contestability. In this way, HuNI acknowledges 'difference' in the representation and organisation of knowledge and underlines the way in which gender (and other 'differences') are socially produced rather than empirical givens.

ARC (Laura Mandell): ARC, the Advanced Research Consortium, is the parent group overseeing and supporting NINES, 18thConnect, MESA, and the forthcoming ReKN and ModNets. These period-specific, online finding aids and scholarly communities work with traditional scholars to encourage digital production: they provide peer review for digital projects, and guidance along the way. ARC supports that work by hosting the web interfaces for these groups, the SOLR server feeding them metadata, providing the Lucene search engine that powers the online finding aids. ARC also coordinates with proprietary companies to import their data into these online finding aids in order to make them into comprehensive research environments. Finally, ARC is attempting to build a sustainability plan for the group by selling BigDIVA, the Big Data Infrastructure Visualization Application (http://www.bigdiva.org).

ARC is not explicitly a feminist infrastructure, but it is noteworthy that it was originally designed by a feminist, Bethany Nowviskie, and is currently directed by a feminist, Laura Mandell. It is possible, however, that the ethic of service and care is a form of stealth feminist infrastructure (Brown, 2008), channeling technological and social structures toward serving and educating the academy. Indeed, this infrastructure has a service focus: the period-specific groups attempt to serve scholars, to educate them about best practices in the peer-review procedure, to work with them in building up projects and creating metadata about those projects. ARC serves the period-specific communities so that they can focus on outreach to traditional scholars. At a recent two-day ARC meeting, including one or two directors from each period-specific group, we spent a full fifth of our meeting time discussing sexism and its impact on our work: the ARC board is made up of feminist men and women attempting to intervene in the way that the academy does its normal business.

CWRC (Susan Brown): The Canadian Writing Research Collaboratory is a feminist project in drag: built on a feminist recovery initiative (Brown et al 2006-15), it was a bid to gain infrastructure funding for several projects in women's writing, but its scope was strategically broadened and generalized to 'writing in and about Canada'. At least in 2009 in Canada, focusing on ostensibly neutral infrastructure was apparently a better strategy for gaining funding than a research grant application focused on women's writing. Hence 'drag': we weren't disguising who we were, but dressing ourselves up in infrastructural (read male) clothing, performing the 'neutral'. As a virtual research environment, CWRC tried to take on board the insights of scholars such as Lucy Suchman that we needed to design for situated, embodied, that is sexed and gendered, subjects. Nevertheless, the project struggled throughout with the divide between form and content, technical and subject expertise, that has largely characterized the division between the digital humanities and feminist scholarship, even as CWRC supports projects that investigate gender, race, nationality, and sexuality in diverse ways.

Infrastructure, collaboration, and credit

We will conclude by tackling the thorny question of credit, awkward though it is to articulate. Infrastructure development's implicit relation to service means that the development of research infrastructure is not considered research *per se* but rather is consigned to the role of research 'support' and is therefore not 'accountable' as a specific research outcome in a range of contexts. To materialize research infrastructure is a process that is clearly distinguished from the reality produced: a covert operation, not a recognized result. And yet the devising and development of new systems might equally be seen as the devising and development of their standards, measures, and meanings and the principles of their provenance in which the restless, transformative, and connective work of infrastructure can be understood as a form of inventiveness and interpretive resourcefulness too.

Collaboration is both highly prized in the digital humanities as a privileged, indeed essential (Price, 2011: 9) form of scholarly practice. Yet, we would contend on the basis of a number of stories arising from our own experience and knowledge of others', it is also fraught for women to the extent that their contributions are more prone than men's to be overlooked or misattributed. We will provide a few largely anonymized examples of having work misattributed, credit or authorship erased (or absorbed into a project identity), and contributions to fields ignored, followed by discussion of how to combat such tendencies.

The panel, we hope, will help to improve understanding of 1) the extent to which even something as apparently neutral or apolitical as infrastructure is imbued with gender and other socio-political considerations; 2) the impact of systemic gender and racial discrimination in a range of infrastructural contexts, notwithstanding the extent to which so many DH practitioners work hard to overcome the biases embedded in our cultures and our discourses; and 3) current and prospective strategies for countering those biases. We will seek to engage the audience throughout this session to include in the panel's discussions a broad range of perspectives on and positions in relation to infrastructure.

Successful infrastructure has the capacity to transform the world in which we already (co-)exist. Digital humanities infrastructure can open up new visions of the world in which we live, and invite contemplation of the different ways in which we might live, and work, in it.

Bibliography

Bowker, G. C. and Starr, S. L. (2000). *Sorting Things Out: Classification and its Consequences*. Cambridge, MIT Press.

Brown, S. (2008). 'Delivery/Service.' In Smith, M. N., et al., *Agora. Techno.Phobia.Philia2: feminist critical inquiry, knowledge building, digital humanities*. Panel at Digital Humanities 2008. Oulu Finland. http://www.rch.uky.edu/docs/Digital-Humanities2008BookOfAbstracts.pdf

Brown, S, Clements, P and Grundy, I. (2006-2015). *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*. Cambridge, Cambridge University Press. http://orlando.cambridge.org

Clement, T. (2012). Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind.' In Hirsch, B. D. (Ed.) Digital Humanities Pedagogy: Practices, Principles and Politics. Cambridge, Open Book Publishers. http://www.openbookpublishers.com/reader/161#page/1/mode/2up

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, **14**(3): 575–99.

Harding, S. G. (1986). *The Science Question in Feminism*. Ithaca, Cornell University Press.

Kraus, K. (2013). Alt-Research for Humanities PhDs. Kari Kraus. http://www.karikraus.com/?p=234

Laboria Cubonix Collective.*Xenofeminism: A Politics for Alienation*. http://www.laboriacuboniks.net/

Mahony, S. and Pierazzo, E. (2015). 8. Teaching Skills or Teaching Methodology?, *Digital Humanities Pedagogy: Practices, Principles and Politics* Cambridge, Open Book Publishers, pp. 215–25.http://www.openbookpublishers.com/reader/161#page/1/mode/2up

McCarty, W. (2005). *Humanities Computing*. New York, Palgrave Macmillan.

Posner, M. (2015). The Radical Potential of the Digital Humanities: The Most Challenging Computing Problem is the Interrogation of Power .' The Impact Blog , London School of Economics. First appeared August 12, 2015. http://blogs.lse.ac.uk/impactofsocialsciences/2015/08/12/the-radical-unrealized-potential-of-digital-humanities/

Price, K. M. (2011). Collaborative work and the conditions for American literary scholarship in a digital age. In Earhart, A. E. and Jewell, A. (Ed.) *The American Literature Scholar in the Digital Age*. Ann Arbor, University of Michgan Presss, pp. 9-27. http://dx.doi.org/10.3998/etlc.9362034.0001.001

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, Cambridge University Press.

Terras, M. (2015). On Changing the Rules of the Digital Humanities. Melissa Terras' Blog. http://melissaterras.blogspot.com/2013/05/on-changing-rules-of-digital-humanities.html

Verhoeven, D. and Burrows, T. (2015). Aggregating Cultural Heritage Data for Research Use: The Humanities Networked Infrastructure (HuNI). In Garoufallou, E. and Hartley, R. J. (Eds.) *Metadata and Semantics Research, Proceedings of the 9th Research Conference*. Springer, Cham, Computer and Information Science 544, pp. 417-423. http://dx.doi.org/10.1007/978-3-319-24129-6_36

The Scholarly Digital Edition: Best Practices, Guidelines, and Peer Evaluation

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Kenneth Price
kprice2@unl.edu
University of Nebraska, USA

Raymond George Siemens
siemens@uvic.ca
University of Victoria, Canada

Hans-Walter Gabler
Hans-Walter.Gabler@anglistik.uni-muenchen.de
Ludwig Maximilian University of Munich, Germany

Fatiha Idmhand
fatihaidmhand@yahoo.es
Université du Littoral Côte d'Opale (Lille Nord de France), France

Thomas Lebarbé
thomas.lebarbe@u-grenoble3.fr
Thomas Lebarbé University Grenoble - Alpes, France

Elena Pierazzo
elena.pierazzo@u-grenoble3.fr
Thomas Lebarbé University Grenoble - Alpes, France

This panel is devoted to two main questions: "How are the function and qualities associated with a scholarly edition changed when it is digital?" and "Is excellence in scholarly editing promoted by guidelines, articulations of best practices, enhanced peer review, or seals of approval?" The panel is intended for members of the scholarly editing community, whom it wishes to engage in dialogue about these matters, along with a wider audience of prospective scholarly editors and members of the DH community who are engaged with questions of peer review and how to foster best practices.

The immediate impetus behind the panel is "Considering the Scholarly Edition in the Digital Age: A White Paper of the Modern Language Association's Committee on Scholarly Editions" published by the MLA in Fall 2015. The primary purpose of the white paper is to define the major elements of the "digital scholarly edition" and explore their significance for the ways editions are read, used, and evaluated. Publication of the white paper, however, raised a number of related questions that we hope to explore through this panel:

- Are the following criteria fundamental to scholarly editing, regardless of school or approach: transparency, accuracy, appropriateness of method, clear and responsible documentation, and the exercise of critical judgment in representing a full account of the textual situation at stake?
- Can a communitysourced edition also be a scholarly edition?
- What is the point of scholarly editions, as we currently understand them, in an era of mass data?
- Editions can be situated within and draw upon a muchlargerscale text archive. How are such editions related to the larger archive?
- How might different traditions of editing be "networked" or brought together by digital means in relation to the same text?
- How do editions relate to largescale textual research? Must their representational precision and refined encoding be lost if aggregated through a mechanism like *HathiTrust* or *TAPAS*? How can their particular contributionthe depths of their insightsbe maintained when incorporated within a broader cultural analysis?
- How far might the criteria for exemplary digital scholarly editions actually also apply to exemplary digital scholarship?
- To the extent that the standards for digital scholarly editing are formalized and programmatic could a largescale digital scholarly corpus be produced largely algorithmically?

Peer Review

The discussion will then turn to the role of seals or hyperpeerreviews of digital scholarly editions in a range of scholarly communities and contexts as represented by the panelists.

A seal such as that awarded by the MLA Committee provides either added weight to peer review for those publishing with a press, or a potential alternative to peer review for editions that are not published by conventional scholarly presses. One would expect that with the shift towards digital editions the number of applications to the Committee would have risen. This, however, has not been the case. Panelist and members of the community of digital editors in the audience will be asked to assess the need for such seals of approval and to consider the adequacy of general guidelines such as the MLAs for editorial work.

Panelists will reflect on questions including the following:

- How do seals or peer review of digital editing projects operate in various contexts?
- Can a seal to something that doesn't reach closure in the usual way? How far along does it need to be?
- If the seal is awarded for one iteration of a project, how do we clarify what was award worthy and what later work is beyond the bounds of the award?

- Might best practices be better served by evaluating projects and their methodology at the beginning rather than the end of their lifecycles, so they have the opportunity to incorporate the results of the review to improve the project?
- If a lack of peer review has sometimes plagued digital scholarship, why have digital editors not done more to avail themselves of this kind of opportunity?
- Is the seal regarded as an additional burdenone more hoop to jump throughpotentially delaying production and unlikely to lead to increased sales of books or increased usage of a digital edition?

Panelists

- **Susan Brown**'s work focuses on borndigital scholarly production and the spectrum from regular scholarly activity to editing, particularly in relation to her leadership of the Canadian Writing Research Collaboratory virtual research environment. She recently joined the MLA CSE.
- **Hans Walter Gabler** will discuss networked editing. The native ground for the scholarly edition in our day, and for the future, is the digital medium. The scholarly edition as digital edition is to be conceived of as processual, relational, and interactive. To open it on and as a web platform is the public beginning of its life in the body of its texts and document images, their interrelation and diachronic stratification; and equally in its body of response texts (*vulgo*: commentary). Comprehensively, the digital scholarly edition should be a dynamic site for incremental enrichment through individual as well as communal research.
- **Fatiha Idmhand** or **Thomas Lebarbé,** as coordinators of CAHIER (Corpus d'Auteurs pour les Humanités Informatisation, Édition, Recherche), one of the consortia comprising TGIRHumaNum (Très Grande Infrastructure de Recherche Humanités Numériques), will address the consortium's support of digital editing projects through an initial application phase, followed by microgranta, training, and workshops. They will also address the role of CAHIER in the improvement of visibility and impact of the editions grouped under its umbrella.
- **Elena Pierazzo** will speak to the issue of standards in relation to her role as recent chair of the Text Encoding Initiative, and address the extent to which, in Europe, in general, where there are so many fragmented traditions and disciplines, the possibility of a single evaluating committee or set of guidelines is not feasible. Thus regular peer review and postpublication reviews are the norm. She will address the question of whether broad editing initiatives such as DiXiT (http://dixit.unikoeln.de) or the Nedimah research group on digital editions (which she cochairs) and the recently announced DARIAHEU initiative Living Sources might usefully contribute to formalizing certain standards.
- **Kenneth Price** will discuss the emergence of the white paper in light of the history of the MLA committee on scholarly editions. In an earlier printonly era the committee endorsed a single model of editing, the "critical edition." Now, at a time of both print and digital editions, and in a changed theoretical environment, the committee (through the white paper) endorses a diversity of additional approaches, including genetic and documentary editing. He will also discuss the challenges of awarding a seal to digital project that may well be later amended.
- **Ray Siemens**, as principal coauthor with Julia Flanders, will discuss the history of, impetus for, and basic thrust of the white paper, situating it in relation to earlier, iterative interventions by members of the CSE and our community among them Peter Shillingsburg's "General Principles of Electronic Scholarly Editions" (1993), Charles Faulhaber's "Guidelines for Electronic Scholarly Editions" (1997), the CSE/TEIC volume *Electronic Textual Editing* (2006), and its publication of guidelines and guiding questions for those preparing and evaluating editions in electronic form and more recent considerations, such as social editing.

Bibliography

Burnard, Lou, Katherine O'Brien O'Keeffe, and John Unsworth, eds. (2006) Electronic Textual Editing. New York: Modern Language Association P. http://www.teic.org/About/Archive_new/ETE/.

Canadian Writing Research Collaboratory. http://cwrc.ca

CAHIER « Corpus d'Auteurs pour les Humanités Informatisation, Édition, Recherche » http://cahier.hypotheses.org/leconsortium

Committee on Scholarly Editions. (2015) Considering the Scholarly Edition in the Digital Age: A White Paper of the Modern Language Association's Committee on Scholarly Editions. New York: MLA. Web. https://scholarlyeditions.commons.mla.org/2015/09/02/csewhitepaper/

DARIAHEU (Digital Research Infrastructure for the Arts and the Humanities): https://dariah.eu

Faulhaber, Charles. (1997) "Guidelines for Electronic Scholarly Editions." December 1997. http://sunsite.berkeley.edu/MLA/guidelines.html.

ITN DiXiT (Digital Scholarly Editions Initial Training Network). http://dixit.unikoeln.de

NeDiMAH (Network for Digital Methods in the Arts and Humanities). Scholarly Digital Editions Working Group. http://www.nedimah.eu/workgroups/scholarlydigitaleditions

Shillingsburg, Peter. (1993) "General Principles for Electronic Scholarly Editions." December 1993. http://sunsite.berkeley.edu/MLA/principles.html.

Siemens, Ray, Meagan Timney, Cara Leitch, Corina Koolen, and Alex Garnett, with the ETCL, INKE, and PKP Research Groups. (2012) "Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media." *Literary and Linguistic Computing* 27.4: 445461. http://llc.oxfordjournals.org/content/27/4/445.full

Text Encoding Initiative. http://www.teic.org/index.xml

Are the Digital Humanists Prepared for Open Access to Research Data ?

Vittore Casarosa
casarosa@isti.cnr.it
ISTI-CNR, Italy

Seamus Ross
seamus.ross@utoronto.ca
University of Toronto

Anna Maria Tammaro
annamaria.tammaro@unipr.it
University of Parma

This panel aims to stimulate a debate among the panelists and the audience about the challenges that the new requirements for the dissemination of research results and research data (often referred to as Science 2.0) is posing to researchers, information professionals and their institutions.

Digital technologies continue to reshape the way researchers do research. They are changing not only the way research is conducted, but also the way in which research results are published and disseminated. More and more legislators, funders, publishers, and scholarly research communities require that "research results" include curated and/or raw research data, methods and processes, tools and software, blogs and wikis, in addition to the "traditional" published material such as articles, monographs, conference proceedings, thesis, grey literature, etc..

In addition to that, we have seen in recent years the push (especially from the European Commission) towards the so called "Open Science", where all published material should be Open Access, i.e. freely accessible by anyone interested. The main argument in favor of Open Access is that most of research is being done with public funds, with research results and research data being produced in the public interest, and therefore they should remain publicly available. Of course, Open Access does not prevent commercial exploitation and protection of the research results and the research data, with patents and copyrights.

While there is a general agreement that Open Access is desirable and ultimately should be the main mode of publication, nevertheless its implementation still presents today a number of issues, made even more apparent by the additional requirement of publishing as Open Access also the raw material underlying the research process, generically indicated as "research data". The variety and quantity of (digital) information that needs to be processed, curated, stored, preserved, and made available in Open Access is posing new kinds of challenges to the researchers themselves (the data producers), to the information professionals dealing with that data (the data curators), and to the institutions responsible for the research activities and the data management (such as institutional repositories and libraries).

The issues range from technological issues (access to and interoperability of research data) to institutional and managerial issues (variety of institutional models and data management approaches), from financial issues (setup and maintenance of data infrastructures) to cultural and behavioral issues (reward structures as a necessary component for promoting data access and sharing practices)

Very often research in the Humanities has been done "in isolation", i.e. by a single researcher who gathers "raw data" (sometimes over many years) and based on them produces research results. More recently we often see that Digital Humanists are working in collaborative interdisciplinary team and could be willing to share the raw data, which are the basis for the research results. However, traditional evaluation criteria are an obstacle to this new behavior, as very often only traditional publication are recognized as such. In this context it is understandable the reluctance to publish also the raw data, as they could be the basis for further traditional publications, needed to get the desired recognition.

The panel will try to discuss some of the main issues related to Open Access to research data:

- Scholarly dissemination in the Humanities and in the Sciences
- The economic challenges of Open Access (Gold OA, Green OA)
- Interdisciplinary research and Open Access
- What are "research data" in the Humanities?
- Curation of research data (quality, accessibility, preservation)
- The role of university presses and academic/research libraries
- Publishing monographs or publishing in journals?
- Career issues and academic "reward system"

Each panelists will have about five minutes to present his/her views on some of the issues mentioned above, in order to stimulate a discussion among the panelists, which (hopefully) will trigger also a debate with the audience.

Panel moderator

- Vittore Casarosa (ISTI-CNR, Pisa, Italy)

Panel participants

- Marin Dacos (Centre pour l'édition électronique ouverte – Cléo, France)
- Lorna Hughes (University of Glasgow, UK)
- Maurizio Lana (Università degli Studi del Piemonte Orientale, Vercelli, Italy)
- Claudine Moulin (University of Trier, Germany)
- Seamus Ross (University of Toronto, Canada)
- Anna Maria Tammaro (University of Parma, Italy)

Developing Local Digital Humanities Communities: The Atlanta Studies Network

Brennan Collins
brennan@gsu.edu
Georgia State University, United States of America

Joe Hurley
jhurley@gsu.edu
Georgia State University, United States of America

Sarah Melton
sarah.melton1@gmail.com
Emory University, United States of America

Pete Rorabaugh
prorabaugh@gmail.com
Kennesaw State University, United States of America

Marni Davis
marnidavis@gsu.edu
Georgia State University, United States of America

Michael Page
michael.page@emory.edu
Emory University, United States of America

Ruth Dusseault
rd@ruthdusseault.com
Georgia State University, United States of America

Jay Varner
jayvarner@gmail.com
Emory University, United States of America

Ben Miller
miller@gsu.edu
Georgia State University, United States of America

Robert Bryant
robcbryant@gmail.com
Georgia State University, United States of America

Jeff Glover
jglover@gsu.edu
Georgia State University, United States of America

Robin Wharton
robin.s.wharton@gmail.com
Georgia State University, United States of America

Atlanta might be best known around the world because it was burnt down during the Civil War. While many residents imagine that General Sherman's march to the sea is the reason we have so few historic buildings, the fact is the city has recreated itself repeatedly, knocking down the old and building the new. This trend has led many scholars interested in the city to turn to digital tools to better research and tell the stories of the city.

What started as a series of projects and platforms within the silos of departments and institutions at Georgia State University (GSU) and Emory, has turned into a broadly interdisciplinary network of Digital Humanities scholars and instructors across local institutions. This panel will explore the possibilities that arise when scholars in many disciplines from universities in the same city combine resources and begin to build projects and platforms with a local audience in mind. The panel will begin with an overview of the network, and then, one or two speakers from different institutions and disciplines will discuss each of the following sections. Each topic will be under 10 minutes to allow for discussion during and after each presentation.

Sharing collections

The first part of building our network began through dialogue between Emory and GSU scholars about digitizing archival collections, technical opportunities to produce and share spatial data and maps, and a search for ways to collaborate.

At Emory much of this process was centered on the digitization of a 1928 Atlas of Atlanta and Vicinity, which was the most comprehensive topographic mapping of the city. Over several years Emory students and staff produced a geodatabase with several thematic layers including the street networks, streetcar and rail lines, and building footprints to name a few. The digitization of map collections at Emory has opened new avenues to explore the original content and has amplified the opportunities in the area of digital heritage of Atlanta.

GSU's Planning Atlanta: A New City in the Making, 1930s – 1990s is a digital collection of material related to city planning and urban development in Atlanta. The collection consists of city planning maps and publications, demographic data, photographs depicting planning activities, oral histories, and aerial photographs. This NEH funded project sought to move beyond the traditional digital library model of simply providing digital equivalents of tangible objects. This city planning focused collection provides open access to digitally transformed, dynamic, and engaging content with the goal of enhancing this material for educational and research uses.

During the 1970s, GSU archaeologists, led by Dr. Roy Dickens, conducted excavations associated with the construction of the MARTA rail lines, a major city planning initiative. These materials, over 100,000 artifacts and all the accompanying documentation, represent the single most comprehensive archaeological collection of Atlanta's history. The new efforts to work with this legacy collection are dubbed the Phoenix Collection because, like Atlanta itself, this collection is being reborn. The collection's broader significance stems from the insight it can provide into the development of Atlanta. Because these archaeological materials have accompanying contextual data, they can more easily be connected with other datasets, such as development maps and historical texts, to create a more holistic understanding of the various processes that shaped the development of the city.

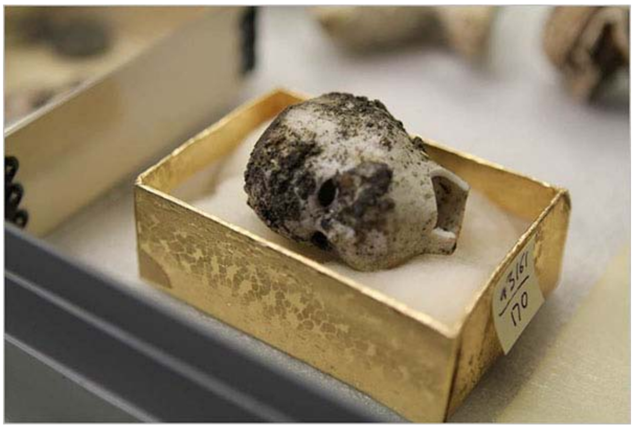


Figure 1. Doll head from Phoenix Collection.

Creating platforms

While digitizing material is crucial for research, making archives more available is only a first step. Instructors and researchers from diverse disciplines found the need to develop platforms for using, sharing, and curating data and archives from different collections.

Discussions about the projects between both institutions and the need to engage the public about Atlanta's history and contemporary scholarship led to the launch of the Atlanta Studies website, an open access, digital publication that any scholar of Atlanta can contribute to. The site hosts many of the projects that make use of the resources and platforms highlighted in this panel and provides a venue for cross-disciplinary work, collaborative institutional work, and public scholarship.

Drawing from the collections of GSU and Emory, the ATLmaps web application combines archival maps, geospatial data visualization, and user contributed multimedia location "pinpoints" to promote investigation into any number of issues about Atlanta. This innovative online platform allows users to layer an increasing number of interdisciplinary data to address the complex issues that cities pose. The project encourages knowledgeable members of the university and local communities to curate data on the site to demonstrate the possibilities for synthesizing material across projects, institutions, disciplines, and data types.

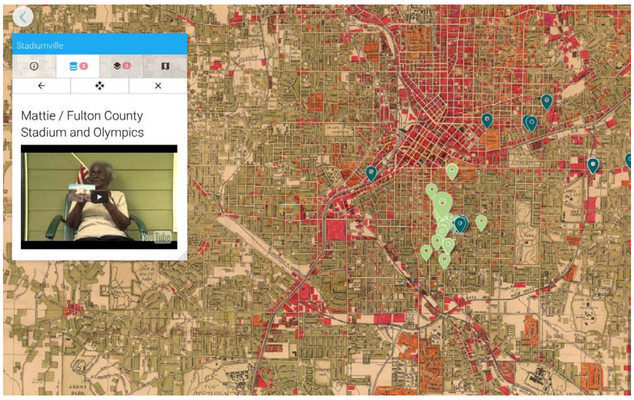


Figure 2. Screenshot of an ATLmaps project.

Emory's Atlanta 3D Explorer is an interactive virtual city circa 1930 built off the 1928 map. The application incorporates the data from the Emory produced geodatabases and historic geocoders in order to create an application that allows users to engage with the content at 3 levels, a map viewer of all atlas pages combined, a 3D city model, and an immersive 3D exploration environment. The third level provides an immersive experience that seeks to recreate street scenes around Atlanta.

In conjunction, a team of students at GSU began building an interactive and immersive 3D version of downtown Atlanta from the late 1920s. The project focuses on city blocks that GSU currently occupies and allows students to explore the history of the area using the Unity gaming engine. Classes in Archaeology, History, and English will eventually help in the research and construction of this 3D environment. Plans are currently being established as to how to integrate these two projects.

Developing a digital pedagogy community

The traditional research and communication skills generally taught in humanities disciplines are crucial to an undergraduate education, but it has become increasingly evident that digital literacy and citizenship need to be integrated into the curriculum. This need runs into the problem that in many humanities departments (where basic academic skills are often taught), digital scholarship is sometimes undervalued, underutilized, and misunderstood. Instructors experienced in or curious about using and teaching digital tools in their classes sometimes lack a community in their own departments or institutions. Atlanta Connected Learning is a collegial network of university faculty, staff, and graduate students from schools in the Atlanta area interested in digital pedagogy, scholarship, and networked collaboration. Since 2012 participants from ATLCL have been instrumental in organizing monthly digital pedagogy meetups, faculty development initiatives, and several smaller academic colloquia.

Developing a local studies pedagogy community

Students learn more if they feel the work they are doing is meaningful and has a real audience. As we have become more aware of ourselves as a network of scholars and instructors, we have begun to connect the Atlanta-based projects we are working on to classrooms. We are currently building a site to connect classes to local projects, resources, and experts that are relevant to the skills and content taught in various courses and disciplines. One example is GSU students in advanced multimodal composition courses are studying specific artifacts from the Phoenix Collection and are creating 3D digital surrogates of these objects through both laser scanning and Structure from Motion techniques. The Phoenix Project provides a visceral means of connecting with the public and bringing to life the past in a way that historical texts alone cannot.

Producing projects

The network that has organically formed over the past five years from the creation and combining of Atlanta-based platforms is now consciously developing projects based on the network that has formed across institutions and disciplines. Scholars, artists, and community leaders have started to contact and join the group understanding the possibility of developing massive projects that are only possible with a great deal of expertise, equipment, and labor.

The first project to intentionally tap into the resources of the network is Unpacking Manuel's Tavern. Students of history, political science, urban planning, sociology, film, architecture, computing and other disciplines will work independently or in the classroom to "unpack" the organic archive that has accumulated over 60 years on the walls of Manuel's Tavern, a bar that has played a significant role in the politics, culture, and history of Atlanta. Faculty and staff from local universities have created a gigapixel map of the tavern walls and a Unity-based 3D scan of the building. Students, directed by instructors from several disciplines, can then choose images they wish to research and compose content for pop-out metadata pages including text, video, interviews and links to other sources from local, national and international archives.

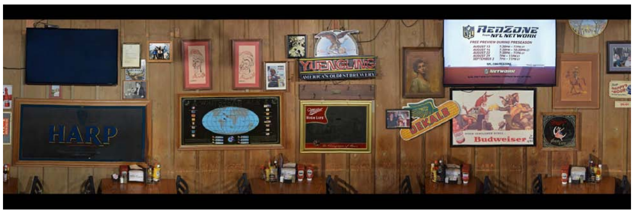


Figure 3. Screenshot of a Manuel's Tavern gigapan.

Digitally Mapping Romantic Literature and Culture

Christopher Donaldson
c.donaldson@bham.ac.uk
University of Birmingham, United Kingdom

Matthew Sangster
m.sangster@bham.ac.uk
University of Birmingham, United Kingdom

Joanna Taylor
j.e.taylor1@lancaster.ac.uk
Lancaster University, United Kingdom

Session overview

This panel session showcases three thematically interlinked papers that report on the work of three digital mapping research projects. Each of these projects engages with specific geospatial methods and technologies to model geo-historical data. Each project, moreover, uses this data to shed new light on the literature and culture of the Romantic Period in Great Britain (c. 1780-1850). One of the projects uses a virtual globe to investigate the effect of movement on writing and authorial practice. Another combines corpus analysis and multimedia GIS to explore readerly practices of reception and spatial re-enactment. The final project engages with curating topographical and cartographic representations and extends the concern with writerly and readerly contextualisation to issues of broader diachronic relevance to literary historiography.

Collectively, these papers examine the links between the physical and discursive geographies of two of the key cultural landscapes of British Romanticism: the bustling metropolitan cityscape of London and the rural upland and coastal terrains of the English Lake District. In discussing the digital mapping of these landscapes, the papers provide new insights about these key cultural locations. In doing so, they extend the pioneering work of previous literary mapping projects, which have focused on the visualisation of authorial experience and on spatial querying of literary texts. Crucially, however, these papers move beyond previous research to explore how linking and juxtaposing different types of data can enhance the study of how places mediate the writing, reading, and critical reception of literary works. In addition, this panel conducts new explorations of the implications of creative mapping for literary criticism, engaging with issues of scale, genre, reception and the limitations and potentials of spatial analysis in humanities contexts.

The first paper, Joanna Taylor's "Path-making and Coleridge's 1802 Tour of the Lake District," uses the virtual globe environment of Google Earth to read the letters, notebooks, and verses related to the poet Samuel Taylor Coleridge's 1802 tour of the western and central Lake District. Taylor demonstrates how using the 3D environment of the virtual globe creates not only new possibilities for assessing Coleridge's journey, but also new possibilities for interpreting a number of canonical Romantic works concerned with walking and the movement of the body through space.

Building on this interest in mobility and path-making, the second paper, Dr Christopher Donaldson's "Deep Mapping the English Lake District," employs a prototype multimedia geographic information system (or "deep map") to explore the geographies underpinning the composition, publication, and reception of the poet William Wordsworth's topographical sonnet series *The River Duddon* (1820). In tracing the cultural influence of this emphatically geo-located lyric sequence, Donaldson's paper explores how digital deep mapping offers new insights about both Wordsworth's practices as a poet and about the material and geographical legacies of his writing.

Following on from this, Dr Matthew Sangster's paper, "Organising Romantic London," presents early conclusions from an ongoing online project that juxtaposes different works which sought to make sense of London by annotating Richard Horwood's "PLAN of the Cities of LONDON and WESTMINSTER the Borough of SOUTHWARK, and PARTS adjoining Shewing every HOUSE." In his paper, Sangster explores the ways in which different genres of source materials attempt to explain the nature of a city of over a million souls, contrasting often-alienated literary responses with the more positive topographical, antiquarian and visual accounts that processes of digital reclamation have brought back to light. He attends to the way that digital methods can help us both to explore the ways in which coherence can be imposed on urban spaces and to expose the limitations of such processes of ordering. In doing so, his paper reflects critically on the roles which digital media can play in preserving and making available archival materials and on the consequences of transforming old media into new digital forms.

Author biographies

Christopher Donaldson is Lecturer in Romanticism at the University of Birmingham and co-investigator on the Leverhulme Trust research project "Geospatial Innovation in the Digital Humanities: A Deep Map of the English Lake District" (2015-2018) and an associate of the European Research Council-funded "Spatial Humanities: Texts, GIS, Places" project (2012-2016). As part of these projects, he is currently completing a monograph entitled *A Literary Atlas of Victorian Lakeland* and a collection entitled *Literary Mapping in the Digital Age* (forthcoming 2016, with Ashgate's Digital Research in the Arts and Humanities series). He sits on the editorial board of the *Journal of Victorian Culture*, and is co-editor of the journal's Digital Forum.

Matthew Sangster teaches at the University of Birmingham and is currently completing the final elements of his first monograph, *Living as an Author in the Romantic Period*. He is Website Editor for the British Association for Romantic Studies and curates the association's blog (http://www.bars.ac.uk/blog/). Between 2008 and 2014, he catalogued the archive of the Royal Literary Fund at the British Library; he is currently developing the catalogue entries into a standalone database. He is working on two new and interrelated projects: one on the development of literary institutions in the eighteenth and nineteenth centuries and the other on the ways in which different genres of works represented London during the Romantic period. Early elements of this project can be seen on http://www.romanticlondon.org.

Joanna Taylor is Research Associate on the Leverhulme Trust research project "Geospatial Innovation in the Digital Humanities: A Deep Map of the English Lake District." She passed her PhD, entitled "Writing spaces: the Coleridge family's agoraphobic poetics," at Keele University in December 2015. She is the British Association of Victorian Studies Newsletter Editor and co-organiser of the ongoing digital project *Placing the Author* (https://placingtheauthor.wordpress.com).

Path-making and Coleridge's 1802 tour of the Lake District

Joanna Taylor

This paper will build on the work undertaken by three digital humanities projects: "Mapping the Lakes" (British Academy, 2007-2008), "Spatial Humanities: Texts, GIS, Places" (European Research Council, 2012-2016) and "Geospatial Innovation in the Digital Humanities" (Leverhulme Trust, 2015-2018). It will read Samuel Taylor Coleridge's (1772-1834) letters, notebooks and poems written on or about his tour of the Lake District in July and August 1802, and suggest how using digital maps to assess this journey opens out possibilities for innovative interpretations of a number of well-known Romantic texts. In particular, the paper will engage with the emergent field of walking studies to explore how physical movement – and its relationship to literary composition – can be productively emphasised using digital mapping tools.

Coleridge spent much of the summer of 1802 traversing the Lake District fells. In part, Coleridge used his pedestrian tour as an escape from his increasingly unhappy marital home, and his considerations of the Lake District landscape assisted him in analysing his personal situation. The "Mapping the Lakes" project charted Coleridge's route, as well as Thomas Gray's important 1769 Lakeland journey, and began exploring subjective elements such as the correlation between elevation and mood. The "Spatial Humanities" project developed from "Mapping the Lakes," and part of that project established a corpus of Lake District texts which now, as part of "Geospatial Innovation," are being analysed using both macro and close reading techniques in order to reveal how text and landscape mutually impacted upon each other. Coleridge's 1802 tour has remained a central text throughout these projects, and this paper will use it as a case study to demonstrate how the digital methods established by this project series can be applied to readings of texts in order to reveal hitherto concealed or unacknowledged elements.

This paper will combine these digital methodologies with approaches from the emergent field of what the cultural geographer Ceri Morgan has termed "walking studies" (2015). Walking studies consider movement through spaces, in ways that may or may not be bipedal, in order to reveal how this movement impacts upon day-to-day lived experiences. As this paper will seek to ascertain, the texts Coleridge produced during and inspired by his 1802 tour – including notebook entries, letters and "Dejection: An Ode" – deserve to be recognised as seminal texts in the walking studies canon.

Coleridge's walking tour enabled him to "command" new territory – both visually and poetically. Whilst climbing up Red Pike, Coleridge considered the relationship between movement and independence: "every man his own path-maker – skip and jump – where rushes grow, a man may go." It is well-known that Coleridge did not restrict himself to the Lake District's already established tourist trails, preferring instead to set off cross-country. As this paper will explore, the diverse terrains he covers are registered in his writing; in particular, his notebooks – which function here as pedestrian journals – indicate his attempts to replicate his walking experiences, both via the way he lays out his entries and his formation of disjointed sentences that attempt to mimic his mode of travel. The paper will aim to demonstrate how these textual embodiments of walking experiences might be recreated or reimagined digitally by interrogating the implications of the 'path' from a multi-disciplinary perspective: literary, geographic and digital.

"Path" is a slippery term. In his famous essay "Walking in the City," Michel de Certeau identifies the path – and, by extension, 'path-mak[ing]' – as one of the crucial modes of operation in constructing space (for de Certeau the imaginative conception of the physical place). He writes that the "intertwined paths" created by footsteps "give their shape to spaces" and "weave places together." Walking, then, is an act of construction in creating the pathways that define spaces. As Paul Cloke has argued, these theories might now productively be applied to non-urban spaces, too. Coleridge's path-making anticipates de Certeau's belief that constructing one's own path is an important means by which to internalise place for imaginative use. Furthermore, Coleridge's accounts of his path-making emphasise the connection de Certeau makes between walking and writing. There is a "rhetoric of walking," de Certeau writes, and this "rhetoric" is constructed through forms and figures in much the same way as a literary work. Furthermore, paths constructed by walking can – indeed, should – be "read" as a literary pursuit.

For de Certeau, the map does not allow for the necessary recognition of the importance of "intertwined paths" to place making. The map displays a set of power relations that inevitably subordinates individual experience. This paper will argue that digital maps offer a way of visualising and re-assessing the role of the path in place making. It does not seem coincidental that the structuring feature of the computer system is also the path: that is, the route to a file, folder, website or other digital destination. Furthermore, programs such as Google Earth draw attention to the connection between computer and physical pathways. Google Earth allows for the creation of "paths" which can digitally trace the physical path created by the walker, and walking websites (such as www.english-lake-district.info) have begun exploiting the potential of Google Earth in facilitating physical walking experiences. This paper will use Google Earth to map such a path of Coleridge's route in the summer of 1802. It will explore the extent to which this kind of digital path can vicariously recreate the physical act of path-making. By combining analysis of this digital path with Coleridge's written accounts of his tour, this paper will uncover a relationship between literary form, terrain and physical movement in ways made uniquely possible by combining literary study with digital techniques.

Deep mapping the English Lake District

Dr Christopher Donaldson

This paper reports on work undertaken as part of the Leverhulme Trust research project "Geospatial Innovation in the Digital Humanities: A Deep Map of the English Lake District" (2015-2018), which is answering the call for scholarship that models the implementation of deep mapping in historical and literary studies research. In this paper, I propose to use a prototype digital deep map to attend to the geographies underpinning the composition, the publication, and – in particular – the reception of William Wordsworth's topographical sonnet series *The River Duddon* (1820), which outlines a journey along one of the Lake District's most important rivers.

Deep mapping involves the accumulation and layering of different kinds of geo-locatable media within a geographic information systems (GIS) environment in order to facilitate investigations of the material, discursive, and imaginative geographies that inform our conception of a location's topography and sense of place. In helping to trace the cultural afterlife of the Duddon sonnets, in particular, the prototype deep map presented here affords insights not only about Wordsworth's reception history, but also about the sense of place communicated in his sonnet series. What is more, the deep map also enhances our understanding of how that sense of place changed and developed through creative re-appropriations during the later nineteenth century.

First published in 1820, Wordsworth's *River Duddon* sonnets have long been recognised as the first sonnet series in English to allow the course of a river to govern the entirety of its design. Wordsworth's sonnets harness the forward-flowing momentum of the lyric sequence to carry the reader on a journey down the river, from its source on the slopes of Wrynose Pass to its estuary. The note that Wordsworth uses as a preface to the sonnets sketches out the itinerary pursued by the series:

> The River Duddon rises upon Wrynose [F]ell, on the confines of Westmorland, Cumberland, and Lancashire; and, serving as a boundary to the two latter counties, for the space of about twenty-five miles, enters the Irish sea, between the isle of Walney and the lordship of Millum [sic].

Each sonnet in the *River Duddon* series moves in succession through this landscape, marking a different spot beside the river and directing the reader's attention to specific landmarks along the way. From Wrynose Pass and Cockley Beck to Seathwaite, Ulpha Kirk, and Swinside, and then finally down to the Duddon sands, Wordsworth's sequence encourages the sense of a continuous journey that carries the reader downstream, at the poet's side, along a single route.

The organic relationship thus developed between the sonnet series and its subject encourages the sense of a continuous journey that conducts the reader from the upland centre of the Lake District to its coastal periphery. Responding to this, readers and literary tourists across the nineteenth century sought to follow Wordsworth's sonnets and to map them onto the landscape of the Duddon Valley. This was accomplished with varying success, for although the *River Duddon* sonnets follow the course of the river (moving in a steady progression), many of the individual sonnets in the series are difficult to locate with precision.

In this paper I will use the deep map to explore these readerly and touristic efforts to locate the geographical sources of Wordsworth's sonnets. My main focus will be to demonstrate how the multimedia GIS environment of the deep map can help us to contextualise the accounts of these readers and tourists alongside other works of literature and visual art that Wordsworth's *River Duddon* sonnets inspired. Specific examples to be featured include the essays and photographs of the Victorian Wordsworth enthusiast Herbert Rix (now held at the Wordsworth Trust in Grasmere), Canon Richard Parkinson's 1843 novel *The Old Church-Clock*, and the sketches of the Duddon valley that R. S. Chattock completed for the Fine Arts Society's illustrated edition of the Duddon sonnets in 1884.

In implementing this approach, this paper indicates the broader ambition of the "Geospatial Innovation" project: to make a major intervention in the application of geographical technologies in the study of the creation and reception of literary works of art. Building on the work and outputs produced as part of the European Research Council-funded "Spatial Humanities: Texts, GIS, Places" project, my aim is to use deep mapping as a means of addressing challenges raised by the application of geographical methods and technologies in literary studies scholarship. A critical issue the project has set out to address is the widely perceived incongruity between the methodologies of geographic information science – with their reliance on precise, quantifiable data – and the kinds of equivocal or "slippery" information with which literary scholars typically engage.

Deep mapping presents a solution to this impediment. The concept of deep mapping emerged out of the psychogeographical experiments of the early French Situationists in the 1960s. More recently, the term has garnered popular interest through the American author William Least Heat-Moon's study *PrairyErth: A Deep Map* (1991), which employs composite, multimedia methodologies to investigate the cultural and historical geographies of Chase County, Kansas. As Least Heat-Moon's *PrairyErth* suggests, deep maps are both topological and relational. Like conventional paper maps they reveal spatial networks that tie locations together. Unlike conventional maps, however, deep maps attempt to record the artefacts, narratives, and memories that underpin those locations and, therefore, shape our understanding of them as places.

Since the appearance of *PrairyErth*, deep mapping has been applied to describe new, exploratory approaches to digital geospatial research. For example, David Bodenhamer and his team at the Virtual Center for Spatial Humanities, Indiana University—Purdue University Indiana, define deep maps as "visual, time-based, and structurally open." "They are," Bodenhamer continues:

> genuinely multi-media and multi-layered. They do not seek authority or objectivity, but involve negotiation between insiders and outsiders, experts and contributors, over what is represented and how. Framed as a conversation and not a statement, deep maps are inherently unstable[.]

Shelley Fishkin, at Stanford University, advocates for a similar conception of the practice:

> Deep maps are palimpsests in that they allow multiple versions of events, of texts, of phenomena (both primary and secondary) to be written over each other – with each version still visible under the layers. They involve mapping, since the form of display – the gateway [. . .] – would be a geographical map that links the text, artifact, phenomenon, or event to the location that produced it, that responded to it, or that is connected with it.

Fiskin's notion that "deep maps are palimpsests" is very salient. There is, after all, an analogical relationship between deep maps and GIS technology: the former is a multi-layered spatial representation; the latter is a tool for integrating layers of geographical data. Exploiting this relationship, the "Geospatial Innovation" project aims to move deep maps from the printed page into a digital environment by harnessing the power and flexibility GIS to store, organise, and visualise a wealth of visual and verbal media. Focussing on the example of Wordsworth's *River Duddon* sonnets, as outlined above, this paper will report on the progress of this research.

Organising Romantic London

Dr Matthew Sangster

In the late eighteenth and early nineteenth centuries, writers producing the kinds of works which we would now call literary were often very sceptical about London. Percy Shelley constructed a considerable part of his poem *Peter Bell the Third* around what for the city is a very unflattering comparison:

> Hell is a city much like London —
>   A populous and a smoky city;
> There are all sorts of people undone,
> And there is little or no fun done;
>   Small justice shown, and still less pity.

While earlier in the eighteenth century, poems such as John Gay's *Trivia* had been able to dwell entirely within the bounds of London, and had found a great deal that was positive to say about the city, late eighteenth and early nineteenth century literary works commonly only dipped into the metropolis, finding the task of systematising either too daunting or not to their tastes. The London poetry and the London poets of the earlier eighteenth century to a large extent fell away, and it was not really until the 1820s, with works like Pierce Egan's *Life in London* and new periodical forms like Charles Lamb's Elia essays, that writers of literary prose moved to take up the nineteenth-century city as a principal subject. Frances Burney's character Evelina circles into London twice to see two quite different sides of metropolitan life, but the city ultimately serves as a place of education through which she must pass rather than as a final destination. William Godwin's Caleb Williams flees into the city (or, more properly, into Southwark), but finds there only temporary and melancholy succour from which he is soon dragged away. William Wordsworth spends a book of the *Prelude* dealing with his residence in London, but he does so in manners which often deflect away or recoil from the city's profusion. When Samuel Taylor Coleridge imagines 'gentle-hearted Charles' from his lime tree bower, he sees him as someone who has "pined | And hunger'd after Nature, many a year, | In the great City pent, winning thy way | With sad yet patient soul, through evil and pain | And strange calamity!" London here is a blockage rather than a solution, a failure of connections rather than a place of fruitful exchange, something that must be overcome rather than a community which sustains.

However, increasingly specialised literary modes of writing were not the only sorts of works which sought to deal with the metropolis. Poetic and fictional accounts wrote back against a wave of topographical and antiquarian accounts which were increasingly seeking to glorify the city. Where poems and novels saw London's size as obstructive and its scale as almost impossible to represent, non-fictional and illustrative accounts of the city commonly perceived its vast scope to be an opportunity. The author of the advertisement to the 1804 volume *Modern London*, probably its publisher Richard Phillips, claims that his book will "exhibit the very soul of the Metropolis [...] Most of the busy haunts of the inhabitants, whether for the gratification of ambition, avarice, or pleasure, have been exactly pourtrayed [sic]; and these views convey at once correct ideas of places which interest from their celebrity, and of scenes which characterize the manners of the people." While Phillips does not claim that London's soul is by any means pure, he does contend that it is graspable – that the city can be characterised and understood, rendered down into a comprehensible and useful written form. For this reason, among others, confident sources like *Modern London* can be enormously useful for accounting for and questioning the gaps left by literary disaffection.

One of the most impressive of the many textual and visual attempts at encompassing London at the end of the eighteenth century was the "PLAN of the Cities of LONDON and WESTMINSTER the Borough of SOUTHWARK, and PARTS adjoining Shewing every HOUSE" produced at considerable financial and personal expense between 1790 and 1799 by the surveyor Richard Horwood. The physical Plan consists of thirty-two printed sheets displaying an area stretching from the middle of Hyde Park in the west to Limehouse in the east and from the southern edge of Islington in the north to the southern fringes of Kennington and Walworth in the south. When assembled, the full Plan is more than thirteen feet across and over seven feet from top to bottom. In its original form, therefore, it is like the city it represents: sprawling, unwieldy and extremely difficult to parse.

The Romantic London website (http://www.romanticlondon.org) seeks to alleviate some of these difficulties through digitally transforming the Plan. The site hosts a detailed tiled version drawn from a source copy in the British Library. This is layered over Google Maps and Open Street Map and can be annotated with text and images (this is done using a straightforward Wordpress plugin). The site makes the Plan generally available, bringing it out of the archive and allowing users dynamic control over its size and scope. It also curates the plan by bringing it into conversation with other means of ordering and organising the metropolis from the late eighteenth and early nineteenth centuries.

As Horwood's Plan and the intertexts which the site pins to it both eloquently demonstrate, London was a prospect both radically and conservatively different from any other contemporary urban or rural environment in Britain. Its quantitative scale mandated qualitative differences which encompassed systems and complexities which had not yet developed, or which were deemed unnecessary, in less extensive urban environments. At the time of writing, the Romantic London website includes a number of means by which contemporary auditors sought to understand the city: plates from Rudolph Ackermann's *Microcosm of London*; plates and text from Richard Phillips' *Modern London*; and text from *Fores's New Guide for Foreigners* and the 1788 edition of *Harris's List of Covent-Garden Ladies*. (I plan to add more curations in the coming months.) Making the geographies implied in these texts explicit through marking them on Horwood's Plan allows us to see the patterns of attention and inattention which they encode. Comparing different curations can bring to light hot and cold spots within the city and can show clearly the varying priorities of different stakeholders in its identity. When put into conversation with less geographically-specific literary works, other kinds of texts can reveal the roles that generic conventions played in shaping London, with other types of writing and visual representations both filling a gap left by literature's retreat from the city and providing key contexts for its eventual reengagement. Modern accounts of London are disproportionately shaped by privileged literary writings; neglected non-fictional accounts can therefore play key roles both in complicating negative literary views of London and recovering the occluded histories of the less fortunate.

The paper will conclude by considering the opportunities and challenges presented by digital remediation for understanding material and media histories. While a digital version of Horwood's Plan is very convenient in many respects, its creation represents an intervention that transforms the Plan's original form in ways that occlude some of its characteristics (size; physicality) and that provide new characteristics unavailable to its original users. Similarly, making texts and images into explicitly geographical systems has considerable implications for their meanings. The digital has the potential to conceal as well as to reveal, and in concluding, I will briefly discuss the ways in which informed curation can serve both to minimise the losses inherent in digital transformation and to maximise their potential for creating new tools and insights which can reify the deeper histories which previous media have encoded but concealed.

Quality Matters: Diversity and the Digital Humanities in 2016

Amy Earhart
aearhart@tamu.edu
Texas A&M University, United States of America

Alex Gil
colibri.alex@gmail.com
Columbia University, United States of America

Roopika Risam
rrisam@salemstate.edu
Salem State University, United States of America

Barbara Bordalejo
barbara.bordalejo@arts.kuleuven.be
KU Leuven, Belgium

Isabel Galina
igalina@unam.mx
Universidad Nacional Autónoma de México (UNAM), Mexico

Lorna Hughes
lorna.m.hughes@gmail.com
University of Glasgow, Scotland, United Kingdom

Melissa Terras
m.terras@ucl.ac.uk
University College London, United Kingdom

Digital humanities has been positioned as an amorphous and fluid concept (Kirschenbaum, 2014), particularized in various disciplines, national contexts and even local environments (Galina Russell, 2014; Fiormonte, 2015; Earhart, 2015; Risam, 2017). Yet intact structures that include the annual digital humanities conference, the various global organizations that form ADHO, and journals published by the various societies represent the field as a coherent body of practice. Ruptures at recent conferences reveal the deceptive constructedness of a coherent digital humanities. What we find, instead, is that digital humanities is deeply divisive and fragmented. Such ruptures were apparent at the 2013 Nebraska conference where the 2014 conference organizers announced an all male keynote lineup;[1] the 2013 conference where the awards committee admitted publicly that they made an error in the graduate student awards selection and had not considered gender, resulting in a skewed awards field;[2] the 2015 conference where the conference program featured an announcement of the all male board of the Frontiers in Digital Humanities journal; or the 2015 conference where the opening set of seven speakers were all male.[3] In light of this recurring theme, our panel examines how digital humanities, as represented by the yearly international conference, is a digital humanities that elides the borders of practice, that masks areas of dissension, and normalizes the field to a particular homogeneous form without contour.

This idea is supported by the conference programs themselves. Scott Weingart's analysis of DH2015 makes clear that "The DH conference systematically underrepresents women and people from parts of the world that are not Europe or North America" (2015). Gender representation in the digital humanities conference is a galling problem. As Weingart's analysis shows, 46% of attendees to DH 2015 were women, but only 34.6% of authors were women (2015). Weingart further notes the data collection challenges of applying quantitative analysis to race, ethnicity, or other categories of identity. Ultimately, though, whether gender, topic, language, nationality or a combination of both, the DH conference has a problem in that the public structure of digital humanities does not represent the depth of work occurring within the field nor its practitioners. The problem is compounded by the treatment of such ruptures as aberrations, temporary mistakes by individuals rather than structural to the makeup of the conference. Furthermore, the organization siloes "diversity" as a committee matter or special interest (GO::DH, MLMC, diversity committee within PC, etc.), rather than a central concern or the universal interest of an organization with global ambitions, an organization whose goals "are to promote and support digital research and teaching across arts and humanities disciplines, drawing together humanists engaged in digital and computer-assisted research, teaching, creation, dissemination, and beyond, in all areas reflected by its diverse membership" (ADHO, 2015). In part this is caused by our insistence on understanding digital humanities as a big tent or a monolithic entity or field, ignoring the ways that digital humanities is practiced in its localized and cultural environment. Whether it is the way that gender, race and ethnicity, and nation are addressed (or not), how funding works in varying countries, the permanence or impermanence of jobs, and other such structural differences, participation in the international DH conference suggests that digital humanities is, in many ways, a living term, ever evolving, ever shifting in response to particular pressures of scholarship, the academy, national and political contexts, and the individual.

A great deal of social science and psychological research has revealed that such struggles are increasingly a problem for organizational groups such as scholarly societies, and that there are productive strategies for engaging diversity questions (Fredette 2015, Plaut 2010). Further, research suggests that a "colorblind" or gender blind response to organizational structures place great strain on the organization and that ignoring issues of difference negative impacts both group dynamics and quality (Apfelbaum, 2012; Dezso and Ross, 2012; Ely, 2001; Nishii, 2013). We will be discussing such findings in relationship to positive, process oriented approaches within the digital humanities conference structure.

Our panel consists of an international group of digital humanities scholars who are invested in a broad understanding of difference and digital humanities. The panel will discuss current tensions within the field and strategies for negotiating the structural challenges to the DH conference and ADHO itself. Of particular concern will be an examination of how the digital humanities conference understands its relationship to the quality of an international scholarship that is inclusive of a broad range of research topics. Panelists will consider topics that include the challenges of negotiating a wide linguistic field, the affordances and limits of quantitative approaches to diversification, practical positive actions that can be taken as part of the work of the program committee, and other pertinent topics. In doing so, they propose an asset - not deficit- model for developing diverse programs that more accurately represent the heterogeneity of digital humanities.

Organization

The panel will open with a 5 minute statement by Amy Earhart, the chair, situating our comments within the above stated context. Each speaker will provide a ten minute statement that will address:

- Barbara Bordalejo, intersectional feminism, the challenges of the majority language and the cultural norm;
- Isabel Galina, peer review, language and perceived quality;
- Alex Gil, translation/multilingualism in the service of trans-cultural equity;
- Lorna Hughes, 'the extra academic professions': the shifting communities of practice in digital humanities, and understanding and recognizing collaboration;
- Roopika Risam, the challenges of quantifying diversity; and
- Melissa Terras, the challenges of managing gender issues in the work of the Program Committee.

After the opening statements, the panel will engage in a discussion with those in attendance. We imagine that our community has much to say about such issues and would like the panel to be a place in which to begin this discussion.

Bibliography

ADHO. (2015). About. http://adho.org/about

Apfelbaum, E., Norton, M. and Sommers, S. (2012).Racial Color Blindness: Emergence, Practice, and Implications. *Current Directions in Psychological Science,* **21**(3): 205-09.

Dezso, C. and Ross, D. (2012). Does Female Representation in Top Management Improve Firm Performance? A Panel Data Investigation. *Strategic Management Journal*, **33**(9): 1072-89.

Earhart, A.E. (2015). Digital Humanities Futures: Conflict, Power, and Public Knowledge. Canadian Society for Digital Humanities (CSDH/SCHN) and the Association for Computing and the Humanities (ACH) Conference, Congress 2015. Ottawa, Canada.

Ely, R. J. and Thomas, D. A. (2001). Cultural Diversity at Work: The Effects of Diversity Perspectives on Work Group Processes and Outcomes. *Administrative Science Quarterly,* **46**(2): 229–73.

Fiormonte, D. (2015). Towards Monocultural (Digital) Humanities? *InfoLet,* http://infolet.it/2015/07/12/monocultural-humanities/

Fredette, C., Bradshaw, P., and Krause, H. (2015). From Diversity to Inclusion: A Multimethod Study of Governing Groups. *Nonprofit and Voluntary Sector Quarterly*, pp. 1–24.

Galina Russell, I. (2014). Geographical and Linguistic Diversity in the Digital Humanities. *Literary and Linguistic Computing,* **29**(4): 307-17.

Kirschenbaum, M. (2014). What Is 'Digital Humanities,' and Why are They Saying Such Terrible Things About It? *differences: A Journal of Feminist Cultural Studies,* **25**(1): 46-63.

Nishii, L. (2013). The Benefits of Climate for Inclusion for Gender-Diverse Groups. *The Academy of Management Journal,* **56**(6): 1754-74.

Plaut, V. C. (2010). Diversity Science: Why and How Difference Makes a difference*Psychological Inquiry,* **21**: 77–99.

Risam, R. (2017). Other Worlds, Other DHs. *DSH: Digital Scholarship in the Humanities,* (forthcoming).

Weingart, S. (2015). Acceptances to Digital Humanities 2015. *Scott Weingart*, http://www.scottbot.net/HIAL/?p=41041.

Notes

[1] Sukanta Chauduri was invited to keynote the conference, the first keynote speaker from the Indian subcontinent, and an important addition to what would have been all Global North male keynotes. At the same time, the lack of a woman keynote speaker suggests the difficulty in balancing a broad set of diversity issues.

[2] The 2013 awards committee publicly admitted their mistake and corrected the error, a highlycommendable action. However, the structural issue remains. We must consider how to develop policies and procedures to avoid replicating such oversights.

[3] The twitter stream of past dh conferences document these incidents and the digital humanities community's reactions. See, for example, Katina Rogers on LLC publication statistics (https://twitter.com/katinalynn/status/358226900616876033), Mia Ridge on the Fortier Prize team's (https://twitter.com/mia_out/status/358340856706646016), Vika Zafrin on the keynote (https://twitter.com/mia_out/status/358340856706646016), Tully Barnett, Deb Verhoeven, and Diane Jakacki on women and Dh 2015 (https://twitter.com/melissaterras/status/621307138070155264).

Literary Concepts: The Past and the Future

Maciej Eder
maciejeder@gmail.com
Pedagogical University of Kraków, Poland

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
Julius-Maximilians-Universität, Würzburg, Germany

Jan Rybicki
jkrybicki@gmail.com
Jagiellonian University, Kraków, Poland

Christof Schöch
christof.schoech@uni-wuerzburg.de
Julius-Maximilians-Universität, Würzburg, Germany

Karina van Dalen-Oskam
karina.van.dalen@huygens.knaw.nl
Huygens Institute, The Hague, Netherlands

Introduction

The methodical discussions on *distant reading* (Moretti, 2013) or *macroanalysis* (Jockers, 2013) have contributed to the awareness of quantitative methods in literary studies, a discipline which many assumed to be the last (and impregnable) stand against the onslaught of quantitative approaches in the Humanities. But actually this late fame is built upon many years of research. Over the last decades, research has been done on style, on aspects of the fictional world like plot and character, on aspects of text groups like genre or historical discourses etc. Most of the research in these fields follows by and large an empirical paradigm. Thus the situation in literary studies is structurally similar to the situation in social science disciplines like psychology or sociology in the past. All of them were using mainly hermeneutical approaches (including some anti-hermeneutical approaches using the same basic methods) until at some point new researchers introduced empirical methods into the field. Interestingly enough, in each case the adoption of empirical methods and the restructuring of the discipline followed its own logic and produced a unique history. In psychology, non empirical methods have been marginalized, and in sociology there is mix of quantitative and qualitative research. This very rough outline does not do justice to the many negotiations and disruptions in the fields, but it is enough to substantiate a point: the adoption of empirical methods in literary studies will probably also be a rather unique story. At the moment, we are right in the middle of it. It seems that empirical methods applied to one or very few texts haven't yet yielded results interesting to larger parts of the fields – with the notable exception of stylometry – while their application to many texts opens up new research possibilities interesting to many. As far as we can see, literary studies is thus in a similar position as some other disciplines in the humanities, they also have their own history of adoption of empirical methods (like linguistics) or are now right in the middle of it.

In this situation, this panel wants to reflect on one aspect of this process: what happens to central concepts of a discipline in these negotiations and disruptions. We want to discuss questions like these:

- Do literary concepts change when DH researchers use them in their (empirical) research?
- If they change, how does this change look like? Can we see commonalities beyond the individual research problems marked by the concepts?
- Using concepts from literary studies in quantitative methods seems to foreground the aspect that many of these concepts are compound and complex notions while empirical research has a pull towards simple, clearly defined notions. If this is true, how do the results of empirical studies translate back to the hermeneutical field?
- Can DH researchers contribute to the ongoing theoretical discussions of these concepts at all and how?
- And finally: Does this tell us something new about the implications DH research has on theoretical concepts from the Humanities?
- We will discuss these question on the basis of five concepts: Genre, Translator, Text, Character and Topic. Each of these concepts has played a role in a sequence of studies using quantitative methods and thus they provide an excellent basis for the theoretical questions we raise.

Text

Maciej Eder

Text as a theoretical concept became obvious for the librarians in ancient Alexandria, who realized that various copies of the same literary work tend to differ, in terms of several textual variants introduced in various papyri (Turner, 2014). This led to the idea of the *archetype* (an ideal representation of the "original") that shines through *imperfect copies*. Since then, philology became an art of textual archeology, where the editor played the role of a demiurge and the author's advocate. In the era of electronic resources, this concept has changed substantially, not only in digital scholarly editions, but also in quantitative text analysis. Traditionally understood as a vehicle of (hypothesized) authorial intention, the text became a sequence of characters bearing some information. A careful philological attitude towards author's words was replaced by: (1) distant reading, which in fact means analysis involving *no reading* at all, (2) focusing on the most frequent words rather than on elaborated rare vocabulary, (3) distorting the original word order by applying the "bag-of-words" model of analysis, (4) high tolerance to textual errors, e.g. imperfect OCR. Such a deep redefinition, however, allows for assessing the history of literature on an unprecedented scale (Moretti, 2013; Jockers, 2013).

Genre

Karina van Dalen-Oskam

Main literary genres such as poetry, play, and novel are divided into different subgenres (mostly called "genres" in daily use). The advances in digital corpora and stylometric tools have broadened the genres that are being studied. Science fiction researchers Nichols et al. in 2014 severely criticized genre theory, stating that it has made 'little tractable progress answering the questions "what is a genre?" and "How is one genre distinguished from others?" in the previous decades'. Nichols et al. discussed the development of a methodology to test genre-related hypotheses in a verifiable and falsifiable way, focusing on science fiction and fantasy. They made use of readers' responses and text analysis to find out how subgenres can actually be distinguished. Elsewhere, Jannidis and Lauer (2014) presented a quantitative analysis in which genre played a role. Using Burrows's Delta on (amongst others) the work of Goethe, they showed that this highly successful method for authorship attribution also works well in distinguishing genres, confirming research by others in this area. This should lead to renewed attention for what a genre is and how genres can be distinguished, thus fulfilling what Willard McCarty wrote in 2010: "Computing machines and scholarly intelligence change each other, recursively".

Character

Fotis Jannidis

The concept of "character" has been intensively researched in the last 20 years. Building on (or fighting against) ideas from the 1960s and 1970s there have been proposals to conceive "character" more as sign or more as a mental model. In digital literary studies there have been attempts to apply these results directly (Zöllner-Weber, 2008), preserving the complexity of the models but basically presupposing a manual analysis of the text. Social network analysis on the other hand provides tools to analyse many and complex social networks and has been applied successfully to plays and novels offering an insight into character constellations (Elson et al., 2010; Trilcke and Fischer, 2015). In this context the question how to identify references to characters in a text, which seemed trivial to traditional literary studies, has been researched in depth emphasizing the difference to named entities in non-fictional texts like news (Elson, 2012: chap 2.4). A recent attempt to address the issue of character types (Bamman et al., 2014) also shows the widening gap between the different modes of modeling literary concepts which will make it difficult to introduce these results into the mainstream discussion on characters.

Translator

Jan Rybicki

Already some of the earliest ventures into the stylometry of translated texts focused on the balance between the translatorial and authorial voices. For a while, it seemed that Venuti's concept of translator invisibility (1995) found vindication in stylometric analysis: in machine-learning classification experiments, translated texts tended to group according to the original author rather than the translator – surprisingly so, since the results were based on translator-generated most frequent words that have little one-on-one correspondence to those in the original (Rybicki, 2012). The application of new visualization methods such as network analysis has allowed to bring out the signal of the translator. The main challenge now is to identify the mechanism of the interaction between the two. Possible directions in this respect include expansion of linguistic features used in analyses, comparison of topic models obtained for originals and their translations, application of text reuse detection tools to translations of the same text and, at the same time, further attempts to bridge the gap between empirical evidence and linguistic theory.

Topic

Christof Schöch

When topic modeling (Blei, 2003) was transferred from information retrieval in expository prose to literary studies, it brought with it the promise of a sophisticated access to the themes in large collections of literary texts. However, "theme" and "topic" are far from identical (Schmidt, 2012; Rhody, 2012): rather, the semantic relation between words grouped in a topic may concern, among others, personnel, setting, narrative motives, or metatextuality as well as (depending on preprocessing) character names, specific registers, or foreign-language terms. Usually, only a minority of topics represent abstract themes. This can partly be explained by the fact that literary texts, unlike expository prose, frequently enact rather than explicitly discuss their most important themes. It also shows that, unlike literary scholars, topic modeling does not distinguish background and foreground information. Finally, the fact that topic modeling is entirely unsupervised and data-driven challenges established hermeneutic strategies. Nevertheless, it seems that established concepts from literary studies such as motive, setting, personnel and metatextuality may help better appreciate the complexity of topic models derived from literary texts. While literary texts challenge any simple understanding of topics, literary studies appears well-equipped to deal with their complexity.

Bibliography

Bamman, D., Underwood, T. and Smith, N. A. (2014). A Bayesian Mixed Effects Model of Literary Character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 370–79.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, **55**(4): 77–84. doi:10.1145/2133806.2133826.

Elson, D., Dames, N. and McKeown, K. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 138–47.

Elson, D. (2012). *Modeling Narrative Discourse*. PhD Thesis, New York City: Columbia University.

Jannidis, F. and Lauer, G. (2014). Burrows's Delta and its use in German literary history. In Erlin, M. and Tatlock, L. (eds), *Distant Readings – Descriptive Turns: Topologies of German Culture in the Long Nineteenth Century*. Rochester: Camden House, pp. 29–54.

Jockers, M. (2013). *Macroanalysis. Digital methods and literary history*. Urbana: University of Illinois Press.

McCarty, W. (2010). Introduction. In McCarty, W. (ed.), *Text and Genre in Reconstruction: Effects of Digitization on Ideas, Behaviours, Products and Institutions*. Cambridge (UK): OpenBook Publishers, pp. 1–11.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Nichols, R., Lynn, J. and Grant Purzycki, B. (2014). Toward a science of science fiction: Applying quantitative methods to genre individuation. *Scientific Study of Literature*, **4**(1): 25–45.

Rhody, L. M. (2012). Topic Modeling and Figurative Language. *Journal of Digital Humanities*, **2**(1). http://bit.ly/1Rmy5n3.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator. In: Oakes, M. P. and Ji, M. (eds), *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*. Amsterdam: John Benjamins, pp. 231–48.

Trilcke, P. and Fischer, F. (2015). Digital Network Analysis of Dramatic Texts. *DH 2015 Abstracts*. http://bit.ly/1M4klcR.

Turner, J. (2014). *Philology: The Forgotten Origins of the Modern Humanities*. Princeton University Press.

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities*, **2**(1). http://bit.ly/1aRoTUo.

Venuti, L. (1995). *The Translator's Invisibility. A History of Translation*. London and New York: Routledge.

Zöllner-Weber, A. (2008). *Noctua literaria – A Computer-Aided Approach for the Formal Description of Literary Characters Using an Ontology*. PhD Thesis, Universität Bielefeld. urn:nbn:de:hbz:361-13097.

Access, Ownership, Protection: The Ethics of Digital Scholarship

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Diane Katherine Jakacki
dkj004@bucknell.edu
Bucknell University, United States of America

James O'Sullivan
josullivan@psu.edu
The Pennsylvania State University, United States of America

Amy Earhart
aearhart@tamu.edu
Texas A and M University, United States of America

Micki Kaufman
micki.kaufman@gmail.com
City University of New York, United States of America,
Modern Language Association of America

Abstract

This panel aims to address the important issue of how we as Digital Humanities scholars negotiate and present the sensitive data (textual, archival, geospatial) that constitutes the core of our analyses. The public facing nature of our work reveals significant challenges that have to do increasingly with access and ethics, and in many cases cause us to reassess how we conduct and disseminate our research. A number of topics pertinent to this issue are addressed in this panel, informed by case studies offered from the panelists' own work. Points of discussion will include, but not be limited to: the negotiation and presentation of sensitive data, access to sources and resources, collaboration, and ownership. In addition to presenting case studies, this panel will incorporate an open dialogue among attendees that addresses these issues across a broader array of research.

Research Ethics in DH

Many of us who work in the Digital Humanities are in some way negotiating, mediating, remediating, and publishing sensitive data. We use the term "sensitive" in a broad way: sensitivity has to do with data that is physically difficult to access because it is privately held, fragile under restrictive copyright, or regards peoples or places that are affected by its use. While our community is opening up new opportunities to digitize, analyze, and share archival materials, the very publicfacing nature of our work reveals significant challenges that have to do increasingly with access and ethics. In many cases, digital methods and approaches make the data upon which we rely even more sensitive. How do we take into account the possibility of risking damage to an artifact in order to digitize it? How do we negotiate rights to data and metadata that has until now been held privately and closely or that involves dozens if not more authors and artists who are still alive, or whose literary and artistic executors have established different parameters for publication? How do we act responsibly when the very publication of a personal work even one that on the surface seems not to impact upon a group of people or a place because of its historical nature may have a profound impact on the lives of those peoples' descendants or the sanctity or environmental protection of those places? The pillars of digital scholarship data visualization and markup, large corpus literary analysis, and geospatial analysis are all complicated profoundly by these questions and sometimes deter us as scholars from working with the very materials we rely on to do our research.

Case Studies

This panel's five contributors work with distinctly different types of data, utilizing a variety of approaches from across the spectrum of digital methods:

Spatial Data

Katie Faull's work with culturalhistorical spatial data in the Susquehanna river watershed has led to a role as mediator between Native American nations and Federal, State and local agencies. Her work regularly raises ethical questions about how data pertaining to sites and landscapes that are carriers of cultural identity and memory for indigenous peoples should be protected from destruction, while at the same time presented to the public as part of important negotiations about conservation. Working with present day Native American nations about the interpretation and conservation of landscapes that are deeply culturally significant, frequently demands delicate negotiations about access and protection and exemplifies the tightrope that many DH researchers must walk. For example, within the realm of Public Humanities, how do we protect indigenous knowledge systems and simultaneously educate the nonindigenous public about those knowledge systems. Her presentation will outline how she and her team of researchers have had to convince archaeologists, for example, that cultural sites can be interpreted within a broader environmental context, thus widening the focus of the cultural historical interpretive lens. At the same time she negotiates with indigenous peoples how to best represent their views of this broader environment. This paper will also discuss how the US Department of the Interior's new landscape conservation initiative (one that is garnering international recognition) may provide a bridge between access, protection and ownership of indigenous cultural memory.

Historically exploited cultural communities

Amy Earhart's work in digital critical race studies has led her to develop community collaborations. However, historical abuses of communities present formidable challenges for those who seek to develop partnerships with vulnerable populations. Earhart will address the challenges that those interested in developing equitable partnerships for collaborative projects might encounter, with particular attention to power dynamics between universities/colleges and such communities using her project The Millican "Riot," 1868 (http://millican.omeka.net) project as a lens through which to discuss how trust and protection might be built into digital projects and how decisions, from project team to technological platform, impact the equitableness of the partnership. In addition, she will discuss strategies for removing control from the academic and the academic institution and, instead, positioning the project within a community or activist site. Finally, she will discuss the use of creative commons licensing including the TK: Traditional Knowledge License and Labeling license (http://www.localcontexts.org) which ensure that historically exploited communities maintain ownership of their material and intellectual property.

Text Analysis

James O'Sullivan's work in computational analytics deals with literary datasets that, while not necessarily restricted, are difficult for peers to replicate. Literary datasets are particularly susceptible to computational approaches, and the new insights that such techniques reveal have the potential to add considerable value to our core disciplines. However, in research contexts where the subject matter is as culturally and socially sensitive as it is intriguing, scholars are presented with an ethical dilemma as far as data is concerned. Many of the works used in macroanalyses are often still under copyright, and so researchers are prohibited from sharing the texts. This restriction precludes our peers from doing two important things: validating our findings, and offering further iterations of our work. Considering the effort that is required in digitising certain datasets, our discipline is fast becoming one where much of the work that claims to be empirically valid cannot in fact be validated. Much of our field's research is conducted on datasets which take the researchers years to acquire and digitize. If datasets are not shared and oftentimes they cannot be replication requires sufficient time and institutional support, and is thus infeasible. As a result, the field has no realistic mechanism by which it can query the validity of methods and interpretations. Should scholars who create datasets hold power over digital artefacts of cultural significance? How can we validate the new insights being offered by scholars in our field? Should we, as scholars, sacrifice access in the name of exploration, or do we need to at least strive for balance between the two?

Intellectual Property

Diane Jakacki's work as a coordinator of digital humanities projects involves collaborating with researchers whose scholarship often entails significant contributions across disciplines and institutions. In addition, the digital nature of this work necessitates the longterm commitment of institutional resources. As this digital scholarship becomes public, questions about intellectual property rights become increasingly complex. Reflecting the nature of DH as a primarily collaborative mode of intellectual work, traditional models of solitary or individual production no longer work. DH collaboration requires teams of investigators, collaborators, data specialists, and research assistants; it often spans years and involves team members across institutions and international borders. The humanistic tradition honors the primacy of the scholar in terms of intellectual property. But as DH methods and forms of labor transform scholarship models, defining primacy becomes ever more complicated. Who is the scholar? Who owns the artifacts that embody DH scholarship? Who makes those determinations? As DH scholars we (rightly) resist any idea that our work is not our own; but do our institutions understand this the same way?

National Security and Public History

Micki Kaufman's work involves the text analysis, data visualization and historical interpretation of the National Security Archive's Kissinger Collection, a carefully curated set of meeting memoranda (memcons) and telephone transcripts (telcons) spanning the 9 years of former US Secretary of State and National Security Advisor Henry A. Kissinger's tenure (1968-1977). Her research into the source material and the methods she has employed to obtain and study it, confront and engage complex issues of copyright and public domain, open access and classification, legality and violence. Through a process of creative deformance of text analytics data (collocation frequency, topic models and other abstractions of text), she uses an aesthetic, visual approach to study patterns and surface complex questions of emotional motivation and behavior, intent and suppression. What ethical features apply when one asks questions about a deeply controversial historical subject's intent and behavior via statistically based text analysis methods, using declassified public domain material from a still classified correspondence, curated by a nonprofit institution and obtained from behind a corporate paywall? From the provenance of the material to the intent of the subject, and from the actions of the declassifying agencies and authorities to the research method, Micki's work examines the ethical underpinnings of the relationship between historian and narrative, setting and subject, word and byte, secret and free.

Bibliography

Beacham, D.(2015). *Indigenous Cultural Landscapes*. http://www.nps.gov/chba/learn/news/indigenous-cultural-landscapes.htm (accessed 3 March 2016).

Borgman, Ch. (2009). The Digital Future is Now: a Call to Action for the Humanities. *Digital Humanities Quarterly*, **3**(4). http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html (accessed 3 March 2016).

Christen, K. (2015). On not Looking: Economies of Visuality in Digital Museums. In Coombes, Annie E and Phillips, Ruth B (eds), *The International Handbooks of Museum Studies: Museum Transformations*. London: Wiley: Blackwell, pp. 365-86.

Christen, K. (2012). Does Information Really Want to be Free? Indigenous Knowledge Systems and the Question of Openness. *International Journal of Communication*, **6**: 2870-93.

Lewis, V. et al. (2015). Building Expertise to Support Digital Scholarship: A Global Perspective. *Council on Library and Information Resources*. http://www.clir.org/pubs/reports/pub168/pub168 (accessed 3 March 2016)

Nowviskie, B. (2011). Where Credit is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship. *Profession*, **13**: 169-81.

Poole, A. H. (2013). Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities. *Digital Humanities Quarterly*. **7**(2) http://digitalhumanities.org:8081/dhq/vol/7/2/000163/000163.html (accessed 3 March 2016).

Werbin, K., C. (2012). The Social Media Contract: On the Paradoxes of Digital Property in this Digital Land. *Journal of Canadian Studies/Revue d'Études Canadiennes*. **46**(2): 245-62.

The National Security Archive. (2015). Most Agencies Falling Short on Mandate for Online Records. http://nsarchive.gwu.edu/NSAEBB/NSAEBB505/ (accessed 3 March 2016)

Digital Data Sharing: Opportunities and Challenges of Opening Research

Natalie Harrower
n.harrower@ria.ie
Digital Repository of Ireland, Ireland

Rebecca Grant
r.grant@ria.ie
Digital Repository of Ireland, Ireland

The data generated in the course of digital humanities research is a valuable resource, with the potential to allow collaborative research and the enhancement of researcher profiles through its publication and re-use. Researchers in the sciences have a longer standing practice of sharing research data, although the pathways to making this data open, discoverable, and interoperable are still being explored. This panel aims to introduce researchers to concepts and best practices in research data management, and to explore the specific challenges and opportunities of research data publication for digital humanities.

Research data is defined as data from any academic discipline which is the subject or product of research. This data may be qualitative or quantitative in nature, and can include documents, spreadsheets, databases, field notebooks, diaries, AV material, transcripts, digital models, algorithms, code, scripts, software applications and other formats. The impetus for research data publication is often closely linked to the principle of Open Access which states that the outputs of publicly funded research should be made openly available and accessible for interrogation by other interested parties. Researchers may be compelled to publish their research data openly due to policies driven by their institution, funder or journal publisher. Currently, the onus is on the researcher to plan their data management and data publication across the lifetime of their research.

Common practice and guidelines have been developed to support researchers in preparing their data for preservation and publication. In many research performing organisations, data curation professionals also provide guidance to support individual researchers. Digital humanities research outputs bring particular challenges, often incorporating or analysing existing copyrighted material. As international policy moves towards mandated preservation and publication of publicly funded research, it is increasingly important that researchers are assisted in identifying, curating and describing their research data, and that Humanities scholars can access training and guidelines relevant to their disciplines and data types.

Internationally, organisations, digital infrastructures and research projects exist to support research data curation activities. The Research Data Alliance (RDA) is an international organisation that draws support and funding from the EU Commission, the National Science Foundation in the United States, and the Australian Government and National Data Services. The RDA aims to accelerate and facilitate research data sharing and exchange, and its work is primarily undertaken through its working groups. Any interested party is welcome to participate in working groups and interest groups, start new working groups, and attend the twice-yearly plenary meetings. RDA Working Groups produce research and tangible outputs (see: https://rd-alliance.org/rda-outputs.html) covering areas relevant to digital humanities including Education and Training on handling of research data, Ethics and Data and Digital Practices in History and Ethnography. Current outputs include recommendations on dynamic data citation, allowing researchers to cite version of changing datasets at certain points in time.

The panel draws together a range of experts in research data management, and will give an overview of key themes in research data and research data management for digital humanities, including identifying and preparing data for deposit, repository infrastructures and the services they provide, and the potential benefits of re-using published research data. The panelists will explore how existing conceptions of research data (often built on the 'hard' sciences) fit with Humanities research, how 'data' may be understood differently depending on domain, and what challenges arise through these differences. A representative from the Research Data Alliance will provide an overview of the RDA's work and its connection to digital humanities research, providing specific examples of relevant outputs to generate discussion from the DH community. This panel will be interactive and welcome engagement from all attendees.

Panel Chair: Dr. Natalie Harrower, Acting Director, Digital Repository of Ireland (Ireland)

As chair, Dr. Harrower will provide an overview of how digital infrastructures can support researchers in preparing and publishing their data. Case studies will include the Digital Repository of Ireland, a national repository for Humanities and Social Science research data which provides services including digital preservation, education and outreach and the creation of persistent identifiers for data citation.

Speaker 2: Rebecca Grant, Digital Archivist, Digital Repository of Ireland (Ireland)

An introduction to concepts of research data in the digital humanities, including accepted definitions of what constitutes research data in a DH context. The specific challenges associated with research data management for DH will be explored.

Speaker 3: Martin Donnelly, Digital Curation Centre (UK)

This paper will present the basics of research data management practice, including the research data lifecycle and the roles and responsibilities of the researcher in managing data.

Speaker 4: Ingrid Dillo, Data Archiving and Networked Services; RDA Technical Advisory Board member (Netherlands)

An introduction to the work of the Research Data Alliance and its relevance to DH researchers. Specific outputs will be presented which have practical applications for the management and reuse of DH research data.

Speaker 5: Orla Murphy, School of English, University College Cork (Ireland)

Opportunities in research data publication for the digital humanities; how digital humanities data publication benefits the data creator, and how DH scholars can reuse research data in their work. A case study showing how published research data can be reused in a humanities context.

Recovering Shared Heritage via Spectral Imaging: Problems, Solutions, Interpretations

Gregory Heyworth
heyworth@olemiss.edu
The Lazarus Project, University of Mississippi, United States of America

Michael Phelps
mphelps@emelibrary.org
Early Manuscripts Electronic Library, Cary, North Carolina

Adrian Wisnicki
awisnicki2@unl.edu
University of Nebraska-Lincoln

Kenneth Boydston
ken@mega-vision.com
MegaVision, Santa Barbara, California

Roger Easton
easton@cis.rit.edu
Rochester Institute of Technology

Chet Adam Van Duzer
chet.van.duzer@gmail.com
The Lazarus Project, University of Mississippi, United States of America

Introduction

This session is planned as a complement to a proposed DH2016 tutorial that will demonstrate multispectral imaging, also organized by the Lazarus Project. The Lazarus Project, based at the University of Mississippi, makes multispectral imaging available to cultural institutions at low cost.

From time to time multispectral imaging garners attention in the press for its success in recovering text and images from damaged manuscripts, maps, and printed books. Yet even among digital humanities professionals, aspects of the process of making multispectral images, from the selection of candidate objects for imaging, to practicalities of setting up for multispectral imaging, the collection of the data, the methods to combine the images to recover the feature(s) of interest, and their interpretation and display, remain obscure. The planned demonstration of multispectral imaging will allow conference attendees to become familiar with the practical aspects of the process, particularly the equipment used, its configuration, and its functioning. The talks in this session will provide a view from a higher vantage point of the practices and results of multispectral imaging. Topics to be addressed include political situations that generate an urgent need for spectral imaging projects, challenges and solutions in the organization of large imaging projects, the different purposes for which spectral images can be used, best practices in image collection, the development of new image processing algorithms to recover text from spectral images, and the use and interpretation of spectral images by scholars.

The talks will be unified by an emphasis on solutions to specific problems in spectral imaging, and it is hoped that this emphasis will facilitate projects by other spectral imaging teams, and will also help scholars in identifying manuscripts and maps that are good candidates for textual recovery through spectral imaging.

The ninety-minute session will consist of six brief papers, as follows:

From Chartres to Timbuktu: Spectral Imaging in a Time of Crisis

Gregory Heyworth

The past hundred years has witnessed greater devastation to the world's collections of ancient manuscripts than any other period since the eleventh century and the First Crusade. Between continued war and climate change, the threat is only increasing. This talk will consider the role of spectral imaging as a response to the political and environmental upheaval. Specifically it will treat various techniques of recovery from spectral and volumetric imaging to x-ray fluorescence on important war-damaged collections in Europe, as well as the logistical challenges of recovery projects on collections in politically unstable areas of the world.

The Sinai Palimpsests Project: The Recovery of Erased Texts in the World's Oldest Library

Michael B. Phelps

St. Catherine's Monastery of the Sinai, Egypt, was built to approximately its form in the mid-6th century and today maintains the world's oldest continually operating library. It holds 4,549 manuscript codices, among which are 160 known palimpsest manuscripts. The palimpsests preserve erased texts in 10 languages that date from the late 4th century to the 12th century. Only three of these 160 palimpsests have ever been the subjects of sustained scholarly study and published. Hence, the palimpsests of Sinai represent a largely unexplored source not only for new texts from antiquity but also for evidence for reconstructing the literary history of the Eastern Mediterranean.

A collaborative project of St. Catherine's Monastery and the Early Manuscripts Electronic Library (EMEL) seeks to recover the erased layers of Sinai palimpsests and make them globally accessible to researchers. Participating scientists are using state-of-the-art spectral imaging and image processing to render the erased texts legible; 23 participating scholars, scattered from Portland to Beirut, are identifying and describing the erased texts based on the image data; and EMEL and the University of California Los Angeles (UCLA) are preparing to host the images and scholarly descriptions online in service to St. Catherine's Monastery. The project represents arguably the most extensive application to date of spectral imaging to cultural heritage.

This talk will survey the methods, innovations, and discoveries of the project. The survey will explore two recurring themes of the project, one technical and the other historical. First, challenges in the implementation of large-scale spectral imaging projects in the cultural heritage arena; and second, the significance of palimpsestation in the transmission history of late antique literature and languages.

The Evolution of Spectral Imaging in the Study of Manuscripts

Adrian S. Wisnicki

The Livingstone Spectral Imaging Project (2010-) is now in its second phase. The project applies spectral imaging to study some of the most damaged surviving manuscripts of David Livingstone (1813-1873), the celebrated Victorian traveler, abolitionist, geographer, and missionary. The first phase of the project (2010-2013) targeted Livingstone's 1871 Field Diary, and sought to use spectral imaging to recover now faded and illegible text that Livingstone had written crosswise over the pages of a single newspaper. The current second phase (2013-2016) centers on Livingstone's 1870 Field Diary, which is of a much more fragmentary nature that the diary previously studied by the project and, in fact, is quite legible under natural light. As a result, in this second phase, the use of spectral imaging has shifted from recovering lost or invisible text to using the imaging technology to explore material aspects of the manuscript, such as after-the-fact-additions (sometimes in other hands) and elements of page topography that can reveal details of the manuscript's passage through time.

Broadly speaking, therefore, the Livingstone Spectral Imaging Project offers a window onto the evolving nature of using spectral imaging technology to study manuscripts during the last five-odd years. The work of the project itself has, in turn, galvanized the development of a larger, but related project, Livingstone Online. Through the collaborative, interdisciplinary nature of its methodologies, the Livingstone Spectral Imaging Project has also established a loosely affiliated network of specialists who have contributed to the development of the two project phases. This paper will reflect on the evolution of this project and give attention to the implications of this evolution, particularly in establishing a collaborative methodology that is empowered by its distributed, at times informal, international character.

Reducing a Challenging Multi-Spectral Imaging Task to Practice

Kenneth Boydston

The multispectral image data set of a cultural heritage object can have a number of uses by disparate parties, and can be in and of itself historically significant. When the object is large, largely degraded, and of significant historical value such as is the late fifteenth-century world map by Henricus Martellus at Yale, acquiring the images that will satisfy known and anticipated demands can be particularly challenging. Time constraints, budget constraints, technological limitations, logistics, handling considerations, risk assessment, personnel, and facilities are among the parameters that can affect the outcome and which should be included in project planning.

Of particular interest are the technological capabilities and limitations. The technology can impact other parameters, and other parameters can impact the technology. While technology will certainly impact the outcome of a multispectral imaging project, it is often the case that technological limitations drive the project—though this should not be the case. In this paper we will suggest that multispectral imaging projects should be driven primarily by the needs of the scholars and conservators who will use the data, and that by focusing on these needs, technological innovations can be created and appropriate technology deployed that will make the difference between mediocrity and success. Examples from the imaging of the Martellus Map demonstrate a few such innovations.

One innovation required by the combination of budget, image spatial resolution, value, size and weight of the object was a large, yet easily portable easel capable of supporting the map and precisely moving both up and down and left and right over a grid of locations. Spatial resolution required that the images be captured in an 11 x 5 grid of tiles and stitched together. Software and hardware protocols were developed to facilitate manual movements of the map. Lasers were integrated to track and maintain focus and camera alignment.

Innovative use of software (MegaVision's PhotoShoot Multispectral Imaging Capture Software) enabled a capture configuration customized for the particular needs of the map. This configuration was planned in advance, and then modified on site as the nature and needs of the map were revealed in preliminary captured samples. The magnitude of the image data set, together with the complexity of processing software (such as PCA), required innovative methods for preparing the imagery for visual appreciation.

Image Processing Techniques for Spectral Images of Historical Objects

Roger L. Easton, Jr., Ph.D.

The history of image processing of manuscripts to enhance or recover erased or damaged text goes back more than 100 years. The first work likely was that of the German physicists Ernst Pringsheim and Otto Gradenwitz in Breslau, who reported in 1895 on the development of an analog method for enhancing the visibility of the undertext of palimpsests by combining pairs of photographic transparencies that were collected under different conditions and processed differently. Their technique was improved by Fr. Raphael Kögel in the 1910s; he used ultraviolet illumination to further enhance the visibility of the undertext in one of the two images, which improved the results. These analog methods were difficult to implement, requiring careful photographic processing and accurate optical alignment to obtain good results. The first use of digital imaging algorithms to recover text from manuscripts may have been by Dr. John F. Benton at the Jet Propulsion Laboratory in the 1970s, who used contrast enhancement and image sharpening tools in the NASA image processing toolkit.

Over the last 30+ years, the capabilities for collecting and processing digital spectral images have improved to the point where it is now possible to have a complete spectral collection and processing system in a single suitcase that may be checked on an airline. The collected images may be processed using a variety of image processing tools, many of which had been developed for use in environmental remote sensing applications and are therefore not available in "general-purpose" image manipulation tools, such as Adobe Photoshop® and the GNU Image Manipulation Program ("GIMP"). Some are available as plugins in ImageJ, but we mostly use the special-purpose package ENVI that is written for remote sensing applications. This talk considers the image processing algorithms that were applied or developed for use to recover text from a wide variety of historical objects.

The algorithms may be loosely divided into two classes: deterministic methods for rendering the image data in pseudocolor, and custom methods that are based on the spectral statistics of the specific leaf. Among the methods in the latter category are principal component analysis (PCA), independent component analysis (ICA), and the minimum noise fraction transform (MNF). In all cases, preprocessing and postprocessing tools are useful. The preprocessing is applied before the deterministic and custom methods and is often necessary to calibrate the images or compensate the image data for differential fading or other problems. The postprocessing tools are used to combine processed images to improve the rendering for scholarly reading.

Examples will be shown of successful processing applied by the author and/or by collaborating team members to a wide variety of manuscripts and other objects in institutions all over the world. Among these are the Archimedes Palimpsest, the Syriac-Galen Palimpsest, the palimpsests in the "New Finds" at St. Catherine's Monastery, the Gruskovà palimpsest in Vienna, and the world map by Henricus Martellus (c. 1491) at Yale.

New Light on Henricus Martellus's World Map at Yale (c. 1491): Multispectral Imaging and Early Renaissance Cartography

Chet Van Duzer

One of the outstanding problems in the history of cartography in the last half century has been that presented by a large world map made by Henricus Martellus in about 1491, which surfaced in the late 1950s and soon came to reside in the Beinecke Library at Yale. The map has long been thought to be one of the most important of the fifteenth century; in particular, there was good evidence that it influenced the thinking of Columbus with regard to his proposed voyage west to reach Asia; it was thought to have influenced Martin Behaim's globe of 1492 and Martin Waldseemüller's famous world map of 1507. But the vast majority of the texts on the map were illegible due to fading and damage, and thus its exact place in Renaissance cartography was impossible to determine.

In this talk I will look at the results of a recent NEH-funded project to make multispectral images of Martellus's map at Yale—at the scholarly use that can be made of the images. These images have rendered almost all of the previously illegible texts on the map legible, and thus have enabled a detailed comparison with Martin Waldseemüller's world map of 1507. This comparison shows that Martellus's map was the source of most of the long descriptive texts on Waldseemüller's map. At the same time, it turns out that the later cartographer chose not to follow Martellus for many of the other details of his map, and thus the comparison generates insights into the workshop practices of an early sixteenth-century cartographer, a subject about which we have very little documentation. In effect the multispectral images enable us to watch Waldseemüller at work, choosing different sources for different categories of information on his map as part of the process of creating a new image of the world.

Diverse Digitalities: Targeted Models for Postcolonial Challenges in the Digital Discourse.

Nirmala Menon
nimmenon@gmail.com
Indian Institute of Technology (IIT), Indore, India

Alex Gil
colibri.alex@gmail.com
Columbia University, NY USA

Rahul Gairola
rgairola@uw.edu
Indian Institute of Technology, Roorkee, India

The digital highway is as yet an exclusive neighborhood-let's just say that there are no traffic jams on there just yet. This disparity is of course not lost on the practitioners of digital humanities and several conversations pointing out this disparity have emerged in the last few years. Whether it is Postcolonial Digital Humanities (DHPoCo), gender algorithms experiments, the articulations of marginalizations spilling over to the digital space have all been part of this discourse. Various projects have also attempted to understand, articulate and bridge the gaps of representations. This panel too is concerned about that gap and discusses different projects that specifically address issues in countries where bandwidth and connectivity is not optimal as in the more advanced nations. How do we harness digital technology so humanities research can be innovative and access to them is not behind a pay wall or a "bandwidth" boundary?

Alex Gil's sx: Archipelagoes project sees to channel the rich and diverse theoretical engagements of the Caribbean with the Digital by providing an innovative two-tiered platform to support digital scholarship in, for and about the region and its diaspora. He proposes a minimum-computing model that can make humanities scholarship in low bandwidth countries longer lasting and easier to access. Alex will also argue for the moral imperative of DH scholars to make that access easy and useful.

Rahul Gairola's & Arnab Datta's project focuses on the theory and practice of electronic education for the under-privileged in India. With their respective humanities and engineering backgrounds, their paper will discuss the imperative of reaching digital technology to the rural population of northern India. As their paper points out, Internet access has still not reached large chunks of the population in India. However, according to Government of India statistics, nearly a billion people will have subscription to mobile phones by the end of the decade. In this context, despite the difficulty of hypertext-based learning in rural areas of India due to lack of internet availability, the solutions can be provided through portable data transfer and convenient access by means of flash memory chips readily available in mobile phones. While many digital archives to date use DVDs and CD-ROMs for archival data storage, no research to date engages alternative data storage and retrieval in rural India archives. State of the art memory technologies support novel technological trends in Postcolonial Digital Humanities – not only in terms of the resources, but also for efficient archival of them. This will make digital literacy in rural India feasible in the immediate future rather than relying on bandwidth sensitive Internet connection.

Nirmala Menon makes a similar case for the dissemination of humanities research and access to it for higher education in India. MAP (Multilingual Academic Publishing) is a new publishing project from Indian Institute of Technology (IIT), India. MAP has two specific features that address the research needs of students and established researchers of humanities in India- 1) it will be an open access platform that will allow access to students of universities across the country and 2) it will publish both original research and translations from and into different languages of India. For this, the project associates are in the process of developing translations software that will be efficient and will aid translations into multiple languages. Nirmala will also argue for the imperative of Digital Humanities to support projects in different postcolonial languages. While there are now DH initiatives in India, a lot of the discourse is still in English even if the projects themselves engage in conversations between different languages. This publishing project ambitiously aims, in the long run, to allow knowledge productions and disseminations in multiple languages.

Together, all three papers address specific problems of the global south and envisage projects that will enable a more diverse global Digital Humanities conversation. The specific examples in these papers are of places that are part of the DH conversation but the geographic locales of the discourse pose challenges that are different from those of EU or US. Each of the projects discusses ways of enabling humanities research and researchers to go from the digital driveway to the highway within the constraints of connectivity and capability.

Small Axe: Archipelagos: A minimal computing model for a digital humanities journal for Caribbean Studies

Alex Gil

The Caribbean is the site of some of the most radical and diverse theoretical and material engagements with the digital. The archipelagos project seeks to channel that activity by providing an innovative two-tiered platform to support digital scholarship in, for, and about the region and its diaspora. Each layer of sx: archipelagos will contribute something new to both Caribbean Studies and to the digital humanities, first via the creation and documentation of a new cost-efficient workflow for the production of text-based scholarly outputs; second, the production of digital humanities project reviews attached to the traditional workflow of book reviews; and third, via the creation and support of a flexible multimodal environment for the production of unique works of digital scholarship that can be ultimately preserved by integrating their components into the university repository. This paper will combine an outline of the specific technological stack needed to run the journal with low resources, and an argument for the moral and practical imperatives to adopt such a model. In brief, I will argue that a minimal computing model can make publication in the humanities longer lasting, easier to access in regions of the world with low-bandwidth and ultimately more transferable to new generations.

Democratic Digitality: Theory and Practice of Electronic Education for the Underprivileged

Rahul K. Gairola & Arnab Datta

This paper is the first co-authored study that deploys questions of gender equity and rural literacy in Literary Studies with electronic memory and communications platforms in the field of Electronics. Our goal herein is to combine our disparate fields to examine, from multiple perspectives, the haunting problems of electronic access to literary/ pedagogical tools in rural India. These problems of access are the consequence of poor infrastructure that demonstrates, as such, the complimentary relationship shared between the humanistic nature of the literary arts and the technical nature of electronics and communications platforms. We would moreover insist that research trends in our individual fields combined allow us to tackle one of South Asia's most pressing issues in the 21st century: digital literacy in rural India. By "digital literacy" here, we do not limit our definition simply to knowledge of knowing how to use digital devices. Rather, we mean the hypertexts, databases, and resources that comprise the soft materials for education and research, and also the required hardware and infrastructure needed to support them. Here, there are significant problems of electronic access to literary/ pedagogical tools in, for example, the requisite data storage and retrieval of digital archives. In principle, these have been realized through the implementation of internet-based hypertexts that facilitate a seamless exchange of knowledge.

However, in the Indian context, the Internet has not yet reached rural areas, and hypertext-based implementation of digital archives is hence not viable. According to the Government of India, nearly a billion users will be subscribing to mobile phones in the coming decade. This is in stark contrast to the growth rate of Internet subscriptions in India, which is projected to be nearly half the growth of mobile phone subscribers. In this context, despite the difficulty of hypertext-based learning in rural areas of India due to lack of internet availability, the solutions can be provided through portable data transfer and convenient access by means of flash memory chips readily available in mobile phones. While many digital archives to date use DVDs and CD-ROMs for archival data storage, no research to date engages alternative data storage and retrieval in rural India archives. State of the art memory technologies support novel technological trends in Postcolonial Digital Humanities – not only in terms of the resources, but also for efficient archival of them. This will make digital literacy in rural India feasible in the immediate future rather than relying on bandwidth sensitive Internet connection. We believe that advanced research and implementation of portable memory devices has the ability to disseminate the resources of the world to the four corners of the earth and improve digital literacy in rural populations.

Web Historiography – A New Challenge for Digital Humanities?

Federico Nanni
federico.nanni8@unibo.it
University of Bologna, Italy

Anat Ben-David
anatbd@openu.ac.il
Open University, Israel

Niels Brügger
nb@dac.au.dk
Aarhus University, Denmark

Meghan Dougherty
mdougherty@luc.edu
Loyola University Chicago, USA

Ian Milligan
i2milligan@uwaterloo.ca
University of Waterloo, Canada

Jane Winters
Jane.Winters@sas.ac.uk
University of London, UK

During the last ten years the use of web archives as primary sources for historical research has attracted the attention of researchers in several different fields. From the web archiving community (Foot et al., 2003; Hockx-Yu, 2014) to Internet studies scholars (Brügger, 2009; Ankerson, 2012), from web scientists (Huurdeman et al., 2013; Hale et al., 2014) to STS researchers (Rogers, 2013; Schafer, 2013), several case studies have been presented. All these different projects have been designed to highlight the potential of born digital sources to offer a new and more complete perspective on our recent past. More recently, traditionally trained and digital historians have been directly engaged in the debate, raising both methodological and theoretical questions (Milligan, 2012; Webster, 2015). However, within Digital Humanities, the web as a source and as an object of study has not been widely debated.

For these reasons, the proposed panel brings together different researchers whose work focuses on the use of born digital materials as historical sources and who have developed, employed, criticized or refused the use of computational/quantitative methods in order to extract useful pieces of information from them.

The purpose of this panel is twofold. First, we aim to discuss how the toolkit available to new generations of historians will necessarily have to combine a vast series of new skills: from accessing and dealing with web archive objects to analyzing networks of hyperlinks, from employing text mining approaches to re-discussing the reliability of a primary source when it is born digital.

Second, our intention is to discuss with the community how this new field of study could be recognized as a relevant example of a digital humanities practice. While the number of born digital sources is increasing rapidly, a broad discussion within the Digital Humanities community is needed, in order to reflect on how to prepare the first generation of digital historians that will work with materials that do not have an analogue counterpart, of which web archives are just one.

Anat Ben-David is a lecturer in the department of Sociology, Political Science and Communication, the Open University of Israel. Her research focuses on Internet geopolitics, web historiography and digital methods for Web research. Ben-David's contribution presents her recent work on the reconstruction of portions of the Web's deleted pasts. In particular, her presentation argues that the use of the Web as a primary source for studying the history of nations is conditioned by the structural ties between sovereignty and the Internet Protocol, and by a temporal proximity between live and archived websites. The argument is illustrated with an archival reconstruction of the history of the top-level domain of former Yugoslavia, .yu, which operated on the Web since 1989 and was discontinued in 2010. The archival reconstruction of a portion of the Web's deleted past serves to assess and conceptualise the Web's limits as an appropriate source for telling its own history.

Niels Brügger is Professor and head of the Centre for Internet Studies as well as of the internet research infrastructure NetLab within the Danish Digital Humanities Lab, Aarhus University, Denmark. His research interests are web historiography, web archiving, and media theory. Within these fields he has published monographs, edited books, and book chapters at international publishers, as well as articles in international peer reviewed journals. He has participated in a number of large research projects, in Denmark and in the UK. Niels Brügger is cofounder and now coordinator of RESAW, a Research Infrastructure for the Study of Archived Web Materials (resaw.eu).

Meghan Dougherty is an assistant professor of digital communication at Loyola University Chicago. Her research focuses on methodological challenges in answering questions about how we move through networks of digital culture to create knowledge, form memory, and make history. The questions that guide her inquiry explore how knowledge production — ranging from identity formation and group membership to scholarly knowledge — is shaped by digital communication infrastructure, and vice versa. Dougherty's contribution to the panel stems from her current book project under contract at University of Toronto Press, *Virtual Digs: Excavating, preserving, archiving, and curating the Web* in which she describes a common field of Web Archaeology drawn together from experiments in Web archiving, digital preservation, and curating that are commonly found in Internet research, digital humanities, and information science. She argues that the nature of scholarly evidence and interpretation must be reconsidered in the new media ecology.

Ian Milligan, an assistant professor of digital and Canadian history at the University of Waterloo (Canada, Ontario) will talk on a specific web archiving project that he has been involved with. His contribution, "WebArchives.ca: Enabling Access to Canadian Political Party Web Archives," explores the development, deployment, and reception of http://webarchives.ca. He argues that since 1996, we have been collecting web archives – now we need to put them to good use. However, accessing and making sense of results requires computational skills. Milligan's case study, a 2005-2015 assemblage of political parties and political interest groups within Canada, should have arguably have been used far more than it had been given pivotal shifts in the Canadian political milieu during the period it studies. These collections were underused, however, due to problems of access restrictions and a lack of technical knowledge. His contribution to this roundtable will then quickly explore various access methods, from content analysis to metadata parsing, using his case study as a reference point throughout.

Federico Nanni is a PhD student in Science, Technology and Society at the Centre for the History of Universities and Science of the University of Bologna and a visiting researcher at the Data and Web Science Group of the University of Mannheim. His research is focused on understanding how to combine methodologies from different fields of study in order to face both the scarcity and the abundance of born digital sources related to the recent history of Italian universities. In particular, he employed oral histories and traditional hermeneutic practices for reconstructing the evolution of the University of Bologna website, which was excluded from the Wayback Machine of the Internet Archive. Later, he applied text mining methods in order to study the collected data. He finally focused his attention on comparing the changes in academic input and output of this institution during the last two decades by analysing the descriptions of the courses presented on the website and the dissertations abstracts available in the digital library.

Jane Winters is Professor of Digital History at the Institute of Historical Research, University of London. Her research interests include the ways in which researchers in the humanities can work with born digital big data, including the archived web. She was Principal Investigator of a big data project funded by the Arts and Humanities Research Council, 'Big UK Domain Data for the Arts and Humanities', which sought to develop a theoretical and methodological framework for the study of web archives over time. Drawing on the lessons of this project she will discuss the challenges faced by researchers wishing to work with the archive(s) of UK web space, and suggest ways in which archiving institutions and researchers can collaborate to overcome them. She will address the fractured and diverse nature of the available web archives - from the open and comprehensive UK Government Web Archive to the dataset derived from the Internet Archive for 1996-2013, from a focused institutional collection such as the Parliamentary Web Archive to the ongoing domain crawl undertaken by the British Library since 2013 - and consider the barriers for researchers in the arts and humanities who choose to use this material, whether as a key source for study or simply as one primary source among many.

Together, these six papers all point towards the growing significance of web archives within contemporary historical practice. In order to encourage the discussion, each speaker will briefly introduce his or her work (5 min), then another speaker will address a general comment (5 min) and following there will be an open discussion (5 min).[1]

The building blocks of the field are in place: what is needed is a discussion within the historical community, but also towards the broader field of digital humanities practice. Our discussion will look to find commonalities and similarities both within our work, but also within the broader DH 2016 conference.

Bibliography

Ankerson, M. S. (2012). *Writing web histories with an eye on the analog past. New Media and Society*, **14**(3): 384-400.

Brügger, N. (2009). *Website history and the website as an object of study. New Media and Society*, **11**(1-2): 115-32.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R.

and Margetts, H. (2014). Mapping the UK webspace: Fifteen years of british universities on the web. *Proceedings of the 2014 ACM conference on Web science*. ACM.

Foot, K., Schneider, S. M., Dougherty, M., Xenos, M. and Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral Web sphere. *Journal of Computer-Mediated Communication*, **8**(4).

Hockx-Yu, H. (2014). Access and scholarly use of web archives. *Alexandria: The Journal of National and International Library and Information Issues*, **25**(1-2): 113-27.

Huurdeman, H. C., Ben-David, A. and Sammar, T. (2013). Sprint methods for web archive research. *Proceedings of the 5th Annual ACM Web Science Conference*, ACM, pp. 182-90.

Milligan, I. (2012). Mining the "Internet Graveyard": Rethinking the Historians' Toolkit. *Journal of the Canadian Historical Association/Revue de la Société historique du Canada*, **23**(2): 21-64.

Schafer, V. and Tuy, B. (2013). *Dans les coulisses de l'internet: RENATER, 20 ans de Technologie, d'Enseignement et de Recherche*. Armand Colin.

Rogers, R. (2013). *Digital methods*. MIT press.

Webster, P. (2015). Will historians of the future be able to study Twitter? http://peterwebster.me/2015/03/06/future-historians-and-twitter/

Notes

[1] A similar format has been already successfully employed at the conference "Web Archives as Scholarly Sources: Issues, Practices and Perspectives", Aarhus, 2015.

Boundary Land: Diversity as a defining feature of the Digital Humanities

Daniel Paul O'Donnell
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

Barbara Bordalejo
barbara.bordalejo@kuleuven.be
KU Leuven

Padmini Murray Ray
p.raymurray@gmail.com
Srishti School for Art, Design and Technology in Bangalore, India

Gimena del Rio
gdelrio.riande@gmail.com
SECRIT (CONICET), University of Buenos Aires, LINHD (UNED)

Elena González-Blanco
egonzalezblanco@flog.uned.es
Spanish Literature and Literary Theory Department, Universidad Nacional de Educación a Distancia UNED (Spain)

> It is normally the case that the objects of scientific inquiry inhabit multiple social worlds, since all science requires intersectional work... The management of this diversity cannot be achieved via a simple pluralism or a laissez-faire solution. The fact that the objects originate in , and continue to inhabit, different worlds reflects the fundamental tension of science: how can findings which incorporate radically different meanings become coherent? (Star and Griesemer, 1989:392)

In the Sociology of Science, objects of scientific enquiry that are common to multiple disciplines or communities are known as "boundary objects" (Star and Griesemer, 1989). As Borgman argues, "these are objects that can facilitate communication, but that also highlight differences between groups" (Borgman, 2007: 153). They have "different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation" (Star and Griesemer, 1989). Handled poorly, they promote "boundary work" - the effort to institutionalize difference between fields (Gieryn, 1983; Lamont and Molnár, 2002: 180); handled well, they are "important interfaces enabling communication across communities" (Lamont and Molnár, 2002: 180).

The theme of this session is the Digital Humanities as a "Boundary Land" - i.e. a locus in which such objects are common. As O'Donnell argues in his paper, this aspect is one of the defining features of contemporary Digital Humanities and an important cause of its recent rapid growth. As the field grows, DH workshops, panels, and journals see increasing work by practitioners trained in more and more traditionally distinct disciplinary traditions: textual scholars, literary critics, historians, New Media specialists, as well as theologians, computer scientists, archaeologists, Cultural Heritage specialists... and geographers, physicists, biologists, and medical professionals.

It is the contention of the speakers of this panel that interpersonal diversity (i.e. diversity along lines such as gender, ethnicity, sexual orientation, language, economic region, etc.) is as an important element of this aspect of DH. The Digital Humanities is not only a place where different disciplines work together (and at times at odds to each other): it is also a place where different people work together and at odds in developing our field. In other words, diversity initiatives in the Digital Humanities are important not only because they let more people into our field, they are important because they change the nature of our field as its practice widens.

The papers in this session each approach the issue from a different perspective. In the first paper, O'Donnell looks at the theoretical background to this understanding of diversity as a component of DH as a boundary discipline, grounding his approach in early work on interdisciplinarity and boundary work. In the second paper, Murray Ray and Bordalejo discuss the ways in which efforts to promote diversity within DH can paradoxically undermine its theoretical importance to the field, before turning to different examples of diversity's intellectual importance. In the third paper, del Rio and González-Blanco examine the institutional and social pressures that promote and hinder dialogue among researchers in developing and developed countries and across linguistic and other boundaries before proposing new approaches in Digital Humanities that go beyond lingüistic diversity focusing on theories such as Sociology of Culture and Education and other reformulations.

All along the Watchtower: an interdisciplinary approach to understanding the importance technical, disciplinary, and interpersonal diversity within the Digital Humanities

Daniel Paul O'Donnell

The increasing paradisciplinarity of Digital Humanities

Perhaps the defining feature of the Digital Humanities as a discipline is its growth (Terras, 2012). Despite some pushback and counter pushback about the precise valence of the field as a discipline (for some more famous recent examples of this growing genre, see Koh, 2015; Fish, 2012; Marche, 2013; Chun, 2013; Grusin, 2013; Jagoda, 2013; a selection of responses to these specific pieces include, among many others, Risam, 2015; Gil, 2015; Liberman, 2012; O'Donnell, 2012; Pannapacker, 2013), digitally inflected work on Humanities problems and material continues to grow.

As its popularity has grown—and, more importantly, as the potential of networked computation as applied to cultural material and questions has become more broadly apparent - it has begun to incorporate practitioners trained in more and more traditionally distinct disciplinary traditions: textual scholars, literary critics, historians, New Media specialists, as well as theologians, computer scientists, archaeologists, Cultural Heritage specialists... and geographers, physicists, biologists, and medical professionals (see Deegan and McCarty 2012 for a detailed discussion of cross disciplinary collaboration in DH).

This growth is interesting for a variety of reasons: as a demonstration of the continuing relevance of the humanities (Davidson, 2011), as a route to new approaches to traditional disciplines (e.g. Ramsay, 2011; Moretti, 2005), and as a method of improving our ability to answer old questions (e.g. Terras, 2006). It also has been interesting for the way it fed back into computer science and other non-humanities domains, for example, through the development of XML and Unicode (O'Donnell 2010).

Above all, however, this growth is interesting because it reflects the increasingly paradisciplinary nature of the domain and its methods. "Humanities Computing", the designation most commonly used before Blackwell's marketing team proposed "Digital Humanities" as an alternative in 2005 (Kirschenbaum, 2010), was far more traditional in approach: beginning with the original work of Roberto Busa in the 1940s, computation in this older form was used to work with relatively traditional objects and questions within relatively traditional humanities domains. As a glance at the tables of contents of journals from this period demonstrates, literary scholars and historians tended to use their computation to do literary and historical work: build concordances and indices, develop statistics, and, later, capture text; Gallery, Archive, Library, and Museum (GLAM) professionals, for their part, computed metadata and built catalogues; Corpus linguists built corpora; and so on. DH, on the other hand, especially in the course of the last decade, has been marked—perhaps defined (O'Donnell, 2012) - by its inter - and cross - disciplinarity: geographers study British Romantic poets' fascination with the Lake District (Cooper et al., 2015); museum curators decipher mathematical texts (Netz and Noel, 2008); literary scholars edit maps, compile archives of things, or analyse Cultural Heritage installations (Foys, 2003; Nelson, 2014; O'Donnell et al., 2012; Hobma, 2014).

Boundary objects and border lines in the sociology of science

In the Sociology of Science, such cross-disciplinary outputs are known as "boundary objects" (Star and Griesemer, 1989). As Borgman argues, "these are objects that can facilitate communication, but that also highlight differences between groups" (Borgman, 2007:153). They have "different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation" (Star and Griesemer, 1989:393). Handled poorly, they promote "boundary work"—the effort to institutionalize difference between fields (Gieryn, 1983; Lamont and Molnár, 2002:180); handled well, they are "important interfaces enabling communication across communities" (Lamont and Molnár, 2002: 180).

The Digital Humanities as Boundaryland

As the Digital Humanities has matured technologically, the question of these Boundary Objects has become an increasingly important, if largely unrecognised, issue among practitioners, users, and critics of DH (Borgman, 2007, chap. 7). As the example of Fish and Liberman (a literary scholar and a computational linguist discussing "distance reading" and "algorithmic criticism") suggests, unacknowledged disciplinary differences about how such objects are understood can result in users talking past each other (see the debate between Fish 2012; Liberman 2012). And indeed a surprisingly large number of other debates in the field can be understood as involving such boundary constructs: complaints that the Digital Humanities are "undertheorised", for example; the "hack vs. yack" debate (i.e. about the relative importance of programming vs. cultural aspects of DH; see Nowviskie, 2014) or even "centre and the periphery" (about the definition of DH through western—or sometimes simply Anglo-American—norms; see Galina, 2013; Fiormonte, 2012).

The rest of this paper looks at the place of globalisation and diversity in light of this understanding of DH as a "boundaryland" - a place in which shared research objects or goals take on different meanings depending on the background of the participant. As the paper will show, interpersonal diversity is in this regard as important to the growth of our field as disciplinary diversity. Research objects, questions, and more importantly, solutions, look very different when lack of funds or inadequate technology prevents the proper preservation of your cultural patrimony and makes participation in collaborations next-to-impossible (see Babalola, 2013). Drawing on the author's experience in establishing and overseeing the first few years of the Special Interest Group, Global Outlook::Digital Humanities, this paper demonstrates the extent to which this openness to difference is in fact an essential feature of both interdisciplinarity and the future growth and development of the Digital Humanities as a discipline that transcends the domains it increasingly incorporates.

Conclusion

Although the concept of the "Boundary Object" is now most commonly discussed in terms of disciplinary difference, the original work on the concept involved the boundary between "Science and Nonscience" or the management of differences in the way in which professional and amateur ornithologists understood the collection of specimens for a museum (Gieryn, 1983). By failing to understand the extent to divisions within the boundaries that coincide with broader cultural, historical, economic, or regional differences amplify existing impediments to the incorporation of the full diversity of our community's experience (Fiormonte, 2012; Galina, 2013; Wernimont, 2013). As Star and Griesemer note:

> When participants in the intersecting worlds create representations together, their different commitments and perceptions are resolved into representations—in the sense that a fuzzy image is resolved by a microscope. This resolution does not mean consensus. Rather, representations, or inscriptions, contain at every stage the traces of multiple viewpoints, translations and incomplete battles... By reaching agreements about methods, different participating worlds establish protocols which go beyond mean trading across unjoined world boundaries. They begin to devise a common coin which makes possible new kinds of joint endeavour (1989, 413).

If You Think You Know the Answer, You Don't Understand the Question

Bárbara Bordalejo and Padmini Ray Murray

Digital Humanities appears to be an open and welcoming field. Indeed, conversations about diversity have been increasingly visible in the digital humanities community. It has been said that the discipline boasts "...a culture that values collaboration, openness, nonhierarchical relations, and agility" and so "might be an instrument for real resistance or reform (Kirschembaum, 59). Notably, this statement by Kirschembaum is also supported by Burdick et al., who state that "...however heterogeneous, the Digital Humanities is unified by its emphasis on making, connecting, interpreting, and collaborating" (24).

These emphases on unity are necessary and vital to ensure that the ideal of a "global" DH establishes itself as a reality in the future. But the widespread belief that these values are at the core of Digital Humanities as a discipline, and that just by virtue of such values it is open and welcoming to all, may prevent us from seeing that the discipline can also fail to meet these standards. And when we discuss the importance of diversity, we need to understand what it is we are talking about: do we mean simply the inclusion of an ever broader collection of participants? Or do we mean that diversity is in some way a crucial intellectual aspect of what we do?

The emphasis on representation and inclusivity has been a main focus of what we might describe as the initial stages in the opening of the Digital Humanities. From the early emphasis on gender balance in the selection of keynote speakers at the ALC/ALLC conference through the early years of DH, to the more recent work of formal committees within ADHO, such as the Special Interest Group Global Outlook :: Digital Humanities (GO::DH) and the the Multilingual/Multicultural Committee, the focus of work thus far has been primarily on discovering and understanding the obstacles that prevent full access to our community and developing strategies to address those obstacles.

There are two dangers to this approach, how ever well meaning and necessary it is.

The first is, of course, that "diversity" is a boundaryless category. By concentrating on one category, or even a few categories, we, almost inevitably end up ignoring others. We focus on gender, but ignore race; or emphasise language, but ignore social class, professional status, or economic/regional disparity.

Thus, when the editorial board of Frontiers in Digital Humanities was announced, it became instantly notorious because all the editors were male. What was less debated was the fact that they were, with one exception, white. Or that they all came from the usually dominant (primarily High Income) countries. Likewise, when Scott Weingart started to write about the acceptance rate of women as first author of papers to the DH conference, it had to do (at least in part) with the relative ease of identifying female and male authors.

In reality, as the example of Frontiers shows, it can be much more difficult to assess other layers of diversity which are not apparent and, therefore, cannot be easily quantified. Even if we imagine that it might be relatively simple to identify native speakers of English, we would remain ignorant of people whose day to day work is carried out in a different language from their native one. Moreover, degrees of bilingualism vary greatly from country to country and culture to culture in such way that for scholars who have English as their second language, their proficiency and ease within it, might be significantly different.

The second danger to this approach is that it trivialises the importance of the category it intends to support. If diversity involves no more than simply ensuring that a wider range of people are present at the table, then questions about the relationship of diversity to quality become, if not entirely reasonable, at least not completely beyond understanding. Perhaps it is possible to become too welcoming--or at the very least to believe that we are somehow watering down the quality of our work by allowing too many participants in simply because they belong to the right demographic.

This paper is about both aspects of the place of diversity within the Digital Humanities.

In the first section, Padmini Ray Murray examines how non-Western cultural concepts and intellectual categories might redefine the digital humanities in terms of methodological frameworks. It is particularly significant that the concept and understanding of what DH is varies in accordance to the cultural context in which it is presented. In other words: although there might be a significant conceptual overlap, one researcher's digital humanities is rarely equal to the one of another. Ray Murray's investigations look at how infrastructural and structural ramifications of working in languages other than English; how notions of the archive can be culturally fluid and how critical making as an intervention is altered by local conditions of production and economics in order to demonstrate that the digital humanities must necessarily be informed by these factors in order to be truly diverse.

In the second section, Bárbara Bordalejo looks at intersectionality and the Digital Humanities. She investigates the combination of factors that might hinder the ability of individual researchers to make themselves more widely known within the DH community. As she shows, background, race, culture, gender, language and ability are all factors whose impact we are just beginning to understand. As she argues, however, this diversity does not need to be a problem. On the contrary, it could be (and it should be) taken advantage of in such way that it challenges and enhances both our research and our community.

A final goal of both parts of this paper is to test the words of Domenico Fiormonte about the importance of social capital in the Digital Humanities: "...it's not enough to have good ideas, work in the Northern [h]emisphere and write them in English: you need good sponsors and authoritative venues." Although it is true that this is part of the problem, nothing should prevent current structures to become part of the solution. Opening authoritative venues can only bring enrichment and new understanding to the DH community.

Spanish Digital Humanities: the construction of a scientific field

Gimena del Rio Riande
Elena González-Blanco García

As Pierre Bourdieu (1975, 19) clearly stated forty years ago: "the scientific field is the locus of a competitive struggle, in which the specific issue at stake is the monopoly of scientific authority, defined inseparably as technical capacity and social power." Universities are strategic spaces for the construction of scientific competences and practices in terms of doxa and habitus (Bourdieu, 1979). The agents that are part of these academic spaces acquire there a socially recognised capacity to speak and act in an authorized and authoritative way in scientific matters. This way, these agents can define and legitimate the definitions they propose for their subjects of study. University education is also a vital experience of utmost importance: university socialization can lead to a deep identity redefinition, with the incorporation of new ways of thinking, communicating and acting. Consequently, altogether with the recognized capacity to legitimate, universities provide the social framework to interpret academic disciplines and communicate specific shared knowledge. (Bernstein, 1990:31), makes it clear when he highlights the distribution of power and principles of control that produce different communication principles unevenly distributed. In his theory, different contexts produce different codes that act selectively on the meanings and realizations.

Undoubtedly, Digital Humanities are nowadays part of the North American and European scientific field. A big offer of postgraduate courses, summer schools, Digital Humanities centers and labs, and scientific journals and websites legitimate the field and its discourse. In this sense, Defining Digital Humanities. A reader (2013), edited by Melissa Terras, Julianne Nyhan and Edward Vanhoutte, can be seen as a text that comes to serve as the last legitimated definition for Digital Humanities. The volume appeared almost ten years after A Companion to Digital Humanities

(2004), edited by Susan Schreibman, Ray Siemens and John Unsworth, and aims to collect the authoritative voices in the field. The book focuses just on voices that have defined the Digital Humanities making use of English as lingua franca and it only considers as authoritative voices a very homogenic group that legitimates the field from universities with a shared Anglo-American perspective and a set of common discourses and practices.

This landscape is very different for the Spanish speaking community. On the one side, European Spanish universities have defined Digital Humanities paying little or no attention to Latin America (González-Blanco, 2013; Spence and González-Blanco, 2014; Rojas Castro, 2013); on the other side, very few definitions have been provided in this side of the world (Galina, 2014; Rio Riande, 2014a, 2014b) or there is a preference for working on a non-defined scientific field that could lead to more open and less philological humanities, more interested in the Social Sciences or Digital Media (Piscitelli 2014). Although there may be many external social, cultural and economic issues that divide the Spanish-speaking Digital Humanities field, this work means to unveal the symbolic violence (Bourdieu, 1991) behind these facts and focus on the characteristics of the institutional spaces in which legitimated discourses and socialization occur as a set of historical and social conditions that explain their particular constitution and nature in Spain and some countries that have started regarding Digital Humanities as a possible (non actual) academic discipline. Knowledge and practices as expressed in the university curricula, but also the set of norms, values and social representations that make each space, can not be fully understood without taking into account the very specific historical, intellectual and institutional factors that have operated and operate in its constitution and legitimate in different ways their discourses.

Regarding the aforementioned, some questions arise: are there possibilities of dialogue in Digital Humanities between developed countries and others with unequal access to technology despite using the same language? Who are the agents that can be part of this dialogue? How do they become part of the scientific field? How much of that symbolic violence comes across in this dialogue? How do social, cultural and historical factors shape the knowledge built at university? The work aims to outline some possible answers to these questions at the time it claims for new approaches in Digital Humanities that go beyond lingüistic diversity focusing on theories such as Sociology of Culture and Education and other reformulations.

Bibliography

Babalola, T. (2013). *The Digital Humanities and Digital Literacy: A Review of Digital Culture in Nigeria.* Digital Studies/Le Champ Numérique.

Bernstein, B. (1990). Poder, educación y conciencia. Sociología de la transmisión cultural.*Esplugues de Llobregat* (Barcelona): El Roure.

Borgman, Ch. L. (2007). Scholarship in the Digital Age: *Information, Infrastructure, and the Internet.* Cambridge, Mass: MIT Press.

Bourdieu, P. (1975). *The specificity of the scientific field and the social conditions of the progress of reason.Sociology of Science Information***14**(6): 19-47.http://ssi.sagepub.com/content/14/6/19.extract(01-11-2015)

-----. (1991). Language and Symbolic Power . Polity Press: Cambridge.

Burdick et al. (2012). *Digital_Humanities. Cambridge: MIT Press.*http://mitpress.mit.edu/sites/default/files/9780262018470_Open_Access_Edition.pdf

Chun, W. H. K. (2013). *The Dark Side of the Digital Humanities – Part 1.*. Center for 21st Century Studies January 9. http://www.c21uwm.com/2013/01/09/the-dark-side-of-the-digital-humanities-part-1/.

Cooper, D., Gregory, I., Bushell S. and Bolton Z. (2015). *Mapping the Lakes.*http://www.lancaster.ac.uk/mappingthelakes/(8-06-2015).

Davidson, C. N. (2011). *Strangers on a Train.* Academe, October. http://www.aaup.org/article/strangers-train#.VDt1_iWCnUQ (01-11-2015).

Deegan, M. and McCarty W. (2012). *Collaborative Research in the Digital Humanities.* Edited by Willard Professor McCarty and Marilyn Professor Deegan. Ashgate.

Earhart, A. (2012). *Can Information be Unfettered? Race and the New Digital Humanities Canon*, Debates in Digital Humanities. U. of Minnesota P. Matthew K. Gold, ed. http://dhdebates.gc.cuny.edu/debates/text/16

English, J. F. (2012). *The Global Future of English Studies*. Chichester, West Sussex, UK; Hoboken, N.J.: John Wiley and Sons.

Fiormonte, D. (2012). *Towards and Cultural Critique of the Digital Humanities.*Controversies around the Digital Humanities, edited by Manfred Thaller, pp. 59–76. Historical Social Research/Historische Sozialforschung 37.1. Köln: Published jointly by QUANTUM [and] Zentrum für Historische Sozialforschung.

-----. (2013) "Seven Points on Multiculturalism" listserve communication May 4th 2013 http://listserv.uleth.ca/pipermail/globaloutlookdh-l/2013-May/000329.html

Fish, S. (2012). *Mind Your P's and B's: The Digital Humanities and Interpretation.*Opinionator. January 23. http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/(01-11-2015)

Foys, M. K (2003).*The Bayeux Tapestry.* Leicester: SDE.

Galina, I. (2013). *Is There Anybody Out There? Building a Global Digital Humanities Community.*Humanidades Digitales. http://humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/ (2-12-2014).

-----. (2014) *Geographical and linguistic diversity in the Digital Humanties*, Literary and Linguistic Computing, **29**(3): 307-316.

Gieryn, T. F. (1983). *Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists.*American Sociological Review, **48**(6): 781–95. doi:10.2307/2095325 .

Gil, A. (2015). *A Non-Peer-Reviewed Review of a Peer-Reviewed*

Essay by Adeline Koh. @elotroalex. April 20. http://elotroalex.webfactional.com/a-non-peer-reviewed-review-of-a-peer-reviewed-essay-by-adeline-koh/ (2-12-2014).

González-Blanco García, E. (2013). *Actualidad de las Humanidades Digitales y un ejemplo de ensamblaje poético en la red*, Cuadernos Hispanoamericanos 761, noviembre 2013.

Grusin, R. (2013). *The Dark Side of the Digital Humanities – Part 2.* Center for 21st Century Studies. January 9. http://www.c21uwm.com/2013/01/09/dark-side-of-the-digital-humanities-part-2/ (2-12-2014).

Hobma, H. (2014). *Digital Killed the Labelling Star: Approaching the Territory-Museum with Mobile Technology.* M. A. Thesis. University of Lethbridge.

Hunt, G. (2015). *There is certainly no gender imbalance in digital humanities!*https://www.siliconrepublic.com/discovery/2015/05/20/theres-certainly-no-gender-imbalance-in-digital-humanities

Jagoda, P. (2013). *The Dark Side of the Digital Humanities – Part 3.*Center for 21st Century Studies. January 9. http://www.c21uwm.com/2013/01/09/the-dark-side-of-the-digital-humanities-part-3/(2-12-2014).

Kirschenbaum, M. G. (2010). *What Is Digital Humanities and What's It Doing in English Departments?*ADE Bulletin, **150**: 1–7.

Koh, A. (2015). *A Letter to the Humanities: DH Will Not Save You.*Hybrid Pedagogy. April 19. http://www.hybridpedagogy.com/journal/a-letter-to-the-humanities-dh-will-not-save-you/(2-12-2014).

Lamont, M. and Molnár V. (2002). *The Study of Boundaries in the Social Sciences.* Annual Review of Sociology **28**(1): 167–95. doi:10.1146/annurev.soc.28.110601.141107.

Liberman, M. (2012). *The "Dance of the P's and B's": Truth or Noise?*, Language Log. January 26. http://languagelog.ldc.upenn.edu/nll/?p=3730 (01-11-2015).

Liu, A.(2012). *Where is Cultural Criticism in Digital Humanities*, Debates in Digital Humanities. U. of Minnesota P. Matthew K. Gold, (Ed) http://dhdebates.gc.cuny.edu/debates/text/20

Marche, S. (2013). *Literature Is Not Data: Against Digital Humanities.* Los Angeles Review of Books . Accessed March 30. http://www.lareviewofbooks.org/article.php?id=1040&fulltext=1 (01-11-2015).

McPherson, T. *Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation*, Debates in Digital Humanities. U. of Minnesota P. Matthew K. Gold, (Ed) http://dhdebates.gc.cuny.edu/debates/text/29

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Nelson, B. (2014). *Investigative Tagging: Modelling the Early Modern Cabinet of Curiosities.* Digital Studies/Le Champ Numérique 0 (0). http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/257 .(01-11-2015).

Netz, R. and Noel W. (2008). Archimedes Codex.. London: Phoenix.

Nowviskie, B. (2014). *On the Origin of "hack" and "yack."*Bethany Nowviskie. January 8. http://nowviskie.org/2014/on-the-origin-of-hack-and-yack/ (01-11-2015).

O'Donnell, D. P. (2010). *Humanities, Not Science, Key to New Web Frontier.* Edmonton Journal, July 21, sec. Technology. http://www2.canada.com/edmontonjournal/news/ideas/story.html?id=e7a48d6c-8a11-4b62-93b3-b48a65e9de74 (01-11-2015).

–––––. (2012). *"There's No Next about It": Stanley Fish, William Pannapacker, and the Digital Humanities as Paradiscipline.* Digital Humanities Now, Editors' Choice, September. http://digitalhumanitiesnow.org/2012/09/editors-choice-theres-no-next-about-it-stanley-fish-william-pannapacker-and-the-digital-humanities-as-paradiscipline-dpod-blog/ (01-11-2015).

O'Donnell, D. P., Karkov C., Rosselli Del Turco, R., Graham, J., Osborn, W., Porter, D., Callieri, M., Dellepiane, M. and Hobma H. (2012). *The Visionary Cross Project : Visionary Cross.* http://visionarycross.org/sample-page/ (21-10-2015).

Pannapacker, W. (2013). *On "The Dark Side of the Digital Humanities."*, The Chronicle of Higher Education. The Conversation. January 5. http://chronicle.com/blogs/conversation/2013/01/05/on-the-dark-side-of-the-digital-humanities/.

Piscitelli, A. (2014). *Cómo definir a las humanidades digitales o cómo no definirlas*:https://media.upv.es/player/?autoplay=true&id=cd36abe4-0f6b-6b4e-9b0a-c4a0d7878203(01-11-2015)

Ramsay, S. (2011). *Who's In and Who's Out* http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out/

–––––. (2011). *Reading Machines: Toward an Algorithmic Criticism.* 1st Edition. University of Illinois Press.

Rio Riande, G. del (2014a). *¿De qué hablamos cuando hablamos de Humanidades Digitales?* en Abstracts de las Primeras Jornadas Nacionales de Humanidades Digitales: Culturas, Tecnologías, Saberes, pp. 17-19 de noviembre de 2014:http://www.aacademica.com/jornadasaahd/toc/6?abstracts

–––––. (2014b). ¿De qué hablamos cuando hablamos de Humanidades Digitales II? http://blogs.unlp.edu.ar/didacticaytic/2015/05/04/de-que-hablamos-cuando-hablamos-de-humanidades-digitales/ (01-11-2015)

Risam, R. (2015). *Revise and Resubmit: An Unsolicited Peer Review. Roopika Risam.* April 20. http://roopikarisam.com/uncategorized/revise-and-resubmit-an-unsolicited-peer-review/ (21-10-2015).

Rojas Castro, A. (2013). *El mapa y el territorio. Una aproximación histórico-bibliográfica a la emergencia de las Humanidades Digitales en España*, *Caracteres*, **2**(2), http://revistacaracteres.net/revista/vol2n2noviembre2013/el-mapa-y-el-territorio/ (01-11-2015)

Schreibman, S., Siemens R. S. and Unsworth J. (eds.) (2004). *A Companion to Digital Humanities.* Oxford: Blackwell. http://www.digitalhumanities.org/companion/ (01-11-2015)

Spence, P. and González-Blanco E. (2014). *A historical perspective on the digital humanities in Spain* en *The Status Quo of Digital Humanities in Spain*, H-Soz-Kult, http://www.hsozkult.de/debate/id/diskussionen-2449 (01-11-2015).

Star, S. L., and Griesemer J. R. (1989). *Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39.Social Studies of Science* **19**(3): 387–420. doi:10.1177/030631289019003001.

Terras, M., Nyhan J. and Vanhoutte E. (eds.) (2013). *Defining Digital Humanities. A reader*. London: Ashgate.

Terras, M. (2006). *Image to Interpretation: An Intelligent System to Aid Historians in Reading the Vindolanda Texts*. 1 edition. Oxford ; New York: Oxford University Press.

–––––. (2012).*Quantifying Digital Humanities.*Melissa Terras' Blog. January 20. http://melissaterras.blogspot.ca/2012/01/infographic-quanitifying-digital.html(21-10-2015).

Weingart, S. (2015a). *Acceptances to Digital Humanities 2015 (Part 1)* http://www.scottbot.net/HIAL/?p=41327 (June 24th 2015).

–––––. (2015b). *Acceptances to Digital Humanities 2015 (Part 2)* http://www.scottbot.net/HIAL/?p=41347 (June 25th 2015).

–––––. (2015c). *Acceptances to Digital Humanities 2015 (Part 3)* http://www.scottbot.net/HIAL/?p=41355 (June 27th 2015).

–––––. (2015d). *Acceptances to Digital Humanities 2015 (Part 4)* http://www.scottbot.net/HIAL/?p=41375 (June 28th 2015).

–––––. (2015e). *What's Counted Counts* http://www.scottbot.net/HIAL/?p=41425 (July 31st 2015).

Wernimont, J. (2013). *Whence Feminism? Assessing Feminist Interventions in Digital Literary Archives.* Digital Humanities Quarterly 7(1).

Intersectional Scholarship in Electronic Literature and Digital Humanities

Élika Ortega
elikaortega@ku.edu
University of Kansas, United States of America

James O'Sullivan
josullivan.c@gmail.com
Pennsylvania State University

Dene Grigar
dgrigar@me.com
Washington State University Vancouver

Introduction

This panel examines some of the many shared issues between Digital Humanities (DH) practices and theories with those in the study of Electronic Literature (E-Lit). Until recently, the fields of DH and E-Lit, though intimately related, have intersected only to a certain extent. The historical development of each of the fields––broadly speaking, DH originating from the earlier humanities computing; E-Lit work from experimental poetics and digital media––might be the reason why these two fields have not engaged in sustained communication. A few events have started to revert this tendency. Within the Alliance of Digital Humanities Organizations (ADHO), the approval of the Special Interest Group on Audiovisual Data in Digital Humanities (SIG AVinDH) is an important milestone fostering the exchange of "knowledge, expertise, methods and tools by scholars who make use of audiovisual data types that can convey a certain level of narrativity: spoken audio, video and/or (moving) images" (AVinDH SIG @DH2015 in Sydney). Similarly, constant collaboration between the Digital Humanities Summer Institute (DHSI) and the Electronic Literature Organization (ELO) in the last few years has lead to deeper collaborations between scholars bridging both fields of knowledge.

In previous ADHO conferences, though on occasion E-Lit has been part of the schedule, it remains a fact that there is room for a timely intervention not only signaling where E-Lit and DH intersect, but also pointing out where E-Lit specific insights are capable to illuminate instrumental approaches to DH theory and practice. E-Lit work has much to offer to DH. An awareness of the expressiveness and historicity of the digital medium that compliments its instrumental and innovative dimension. Dealing with multiple expressive codes besides language, and through the leadership of the ELO, E-Lit work has been at the forefront of inclusivity, diversity, and multilingualism. Issues that continue to be discussed in DH venues.

Specifically, the papers in this panel will focus on the materiality of electronic literature as it illuminates the ongoing debates of print-digital media and changing reading and writing practices; the applicability to works of E-Lit of forms of quantitative criticism that had been used exclusively for print literature in DH; and finally, the best practices for collecting and archiving electronic literature as it affects GLAM and what it can teach us about all, or most, DH work subject to obsolescence. We hope that this panel will foster further explorations and collaborations between the two fields of study.

"Print" Works of Electronic Literature and Scholarly DH Narratives

Élika Ortega

These paper starts with the question what can DH practitioners learn from material composition of electronic literature works in order to improve the outreach of DH projects? This work focuses specifically on the material architecture of "print" works of electronic literature, which I understand as those that depend or heavily hinge on print materials for their digital configuration. Drawing from this corpus, I propose to take E-Lit works as models to explore and address the effect of digital media as it has modified––and continues to propose a modification of––reading and writing practices as well as modes of abstracting, encoding, and communicating information, all of which are common in DH pedagogical and research praxis.

First, I examine a handful of creative print works of electronic literature including Stephanie Strickland's et al *Vniverse* (2011), Amaranth Borsuk and Brad Bouse's *Between Page and Screen* (2012), Nick Montfort's *#!* (2014), and Jacob Garbe and Aaron Reed's *Ice-Bound* (2015). I argue that the architecture of these works requires specific infrastructural conditions and the unfolding of several practices or protocols for their reading that pose challenges not only for preservation and archives professionals, but for the average reader as well. Strickland's , Borsuk and Brad Bouse's , Montfort's, and Garbe Reed's works are so uniquely imagined and crafted that they seem to embed within them the specific critical framework to be theorized. In that sense, they too demand a tailored reading: a look into how their text is media and how their media is text. These kinds of compositions, I argue, have much to teach DH practitioners and students about the expressiveness of the digital medium, the way electronic and print media reciprocally inform, shape, and inflect each other. Further, they allow us to study the modification of the practices and protocols associated with each of its material components: print objects are not self-sufficient and translatable outside of the digital realm, computational devices are rendered useless without input from print materials. Even when these works are highly experimental, the specific conditions of each one constitute a laboratory to investigate contemporary changing reading and writing practices.

Further, I argue that in their radical specificity, creative works like these that rely on various print or digital media for their poetic, material, or narrative construction provide models for multimodal design that is both highly desirable and commonly found in DH scholarly outputs. The organization of information––whether poetic, narrative, or scholarly––is certainly responding to the same moving-target media landscape and, thus, bringing them all together offers an opportunity to observe how material composition is approached in creative works in ways that can be extrapolated to scholarly works. To that end, I extract a handful of compositional strategies from the works mentioned above that can be translated to a scholarly realm. Among them, we can find the use of augmented reality, the design of a script that takes the reader from one medium to another, the *extension* and complementation of information through different expressive languages, and even the performance of reading as an interpretive act. Under this light, aside from being great examples of poetic and computational creativity, these works of electronic literature are capable of illuminating the ongoing debates on the future of the book, and the place of the monograph in academic careers. Crucially, these works also signal the need for training in alternative reading and writing capacities that sits at the center of DH instruction, project development, and outreach. Taken as models of composition in the media ecology in which DH work is currently carried out, these works offer avenues for communication between the two fields not only on a conceptual level, but also as a methods of argumentation and interpretation.

Quantifying the Evolution of Electronic Literature with Zeta

James O'Sullivan

Introduction

This paper seeks to determine to extent to which electronic literature (understood as born-digital literature with an inherent computational aesthetic) has evolved, by analysing the language used to describe the works included in both *Volume I* and *II* of the Electronic Literature Organization's *Electronic Literature Collection*. These anthologies include descriptions of the works by their respective editors, as well as by the contributing authors. Outlining the various technical and literary characteristics reflected in each work, these descriptions provide a unique opportunity to determine the aesthetic qualities of the canon, as depicted by some of the field's most prominent practitioners. Furthermore, the considerable time between the publication of the collections, released in 2006 and 2011 respectively, is such that they provide a useful sample when examining how the electronic literary movement has developed throughout the contemporary era. Using a method typically reserved for print literature, this paper applies a macro-analytical approach to the analysis of these descriptions in an effort to produce some quantitative evidence to support critical interpretations on the evolution of electronic literature. Research of this sort has never been conducted in this field, and thus, not only does this paper develop our understanding of the literary movement in question, but it also breaks new ground in the application of specific computational methods to born-digital artistry.

Methodology & Results

For the purposes of this study, Craig's Zeta was selected as the method best suited to identifying trends in the manner by which electronic literature is described by its creators and curators. A Zeta analysis compares two datasets, producing a set of words distinct to each. In other words, it gives a set of words most likely to occur in Set A, which are unlikely to appear in Set B. In this instance, the analysis compared the editor and author contributions from each volume, providing a list of words which indicate what topics were being prioritised across the collections. Comparing these wordlists, we can see how it is that the focus of the electronic literature community changed over the course of this particular time period. The Zeta analysis was conducted using R, with a text slice length of 2,000, text slice overlap of 1,000, an occurrence of 2, and filter threshold of 0.1. For the purposes of this abstract, the top 25 distinctive words have been displayed (see Table 1), but a more complete set of results will be addressed in the final offering.

Volume I (2006)	Volume II(2011)
Diagrams	video
Exploring	rather
Versions	google
Blue	relation
Clock	see
Galery	herself
Red	pages
Ambient	creative
Awal	water
Corresponding	last
Dhtml	map
Practice	emblems
Earlier	home
Murder	engine
My	production
Version	tree
Line	unknown
Nature	age
Author	beyond
Makes	break
Landscapes	friends
Frome	had
Day	imagery
Poetica	screens
Meditation	starts

Table 1. Top 25 distinctive words used to described the anthologised works

Interpretations

The results of the Zeta analysis provide computational evidence for many of the assumptions that one might make when speculating on this issue. For example, there is a marked shift from static "diagrams" in *Volume I*, to "video", in *Volume II*. The evolution of Web cultures is also apparent, with "dhtml" giving way to "google". Beyond these, somewhat expected, findings, other, more interesting, revelations are also present. In particular, the inclusion of herself" in the later collection would suggest a rise in feminist electronic literature, while words like "exploring" and "poetics" in *Volume I* suggests a field that, in 2006, is still ontologically uncertain. As already noted, a more robust analysis drawn from a more thorough interpretation of the entire set of findings will be offered in this paper. However, from this limited snapshot into this study, it is clear that this approach yields valuable insights into the field, and justifies the use of macro-analysis in the extrapolation of cultural contexts.

Rendering Literature: Methods for Collecting and Archiving Electronic Literature

Dene Grigar

This presentation, entitled "Rendering Literature: Methods for Collecting and Archiving Electronic Literature", focuses on methods developed to document early digital literature, 1986-1995.

This paper builds on research undertaken with *Pathfinders: Documenting the Experience of Early Digital Literature* (with Stuart Moulthrop, scalar.usc.edu/works/pathfinders), a project funded by the National Endowment for the Humanities that developed the methodology for documenting early works of electronic literature (1986-1995) and the print-based book, *Traversals* (forthcoming, The MIT Press, 2016, also with Moulthrop) that provides a critical look at the works themselves. *Rendering Literature* aims to discuss best practices for collecting and archiving electronic literature by libraries, museums and other institutions so that works retain their inherent significant properties, including its cultural context.

Electronic literature is an experimental literary art form that can include a combination of words, images, sound, video, animation, gestures, and movement but always involves code and computation. Referred to often as born digital literature, electronic literature cannot be experienced meaningfully in print and is intended, instead, to be accessed through digital devices. Early work was published on floppy disks, CDs, and DVDs, but the advent of the web made sharing it online with a global audience popular from 1995 onward. The introduction of smart mobile devices in the mid-2000s drove artists to innovate their art for the app environment. To remain accessible to a reading audience, many works of electronic literature have been updated to newer platforms and software iterations––sometimes many times––resulting in numerous versions of a work. In cases of literary art produced as apps, it is not possible to study versions of a work saved on a single device because upgrading to a new version of a work overwrites the previous version completely.

Contributing to the challenge of archiving electronic literature is that many of these works are published as a combination of digital files, accompanying documentation websites, and ephemera. Some, like John McDaid's *Uncle Buddy Phantom Funhouse* (1993) include audio cassettes that are part of the narrative. Others like Judy Malloy's *Uncle Roger Version 3* (1987-8) were packaged in hand-made artists boxes that themselves are works of visual art. Recent works like Erik Loyer's *Breathing Room* (2013) or Amaranth Borsuk and Brad Bouse's *Whispering Galleries* (2014), require additional equipment like a Leap Motion controller. Still others like Jody Zellen's *Urban Rhythms* (2011) exist only as apps. In a word, these works differ widely from traditional digital texts and yet, to date, there are no specific methods used for handling this form of literary art.

A visit in October 2015 to the David M. Rubenstein Rare Book and Manuscript Library

at Duke University to locate data in the Judy Malloy Papers for an article about her database novel, *Uncle Roger*, drove home the need to analyze methods of archiving and preservation undertaken at collections of electronic literature in the U.S. Floppy disks were separated from their artists boxes and unavailable for review, while inserts for the artists boxes were placed in separate folders in a different section of the archival containers. The six versions of the work were not readily distinguishable from one another. Discussions with the librarians and archivists revealed that they too were interested in determining how best to handle such complex problems for works that resist current preservation and archival practices.

Research Significance

Nothing lasts forever. Paper mildews. Sappho's nine books of poetry were burned, leaving but a few extant poems for us to read. Today, poetry and literary forms are increasingly produced in the electronic medium, and the danger facing them is not dampness or fire but the constant innovation of digital technology. Net poetry by artists like Jason Nelson created a mere 10 years ago, for example, is quickly becoming obsolete today because the Apple Corporation decided in 2007 not to support Flash on its iPhones. Despite this problem, digital technologies have fostered the production of so much experimental work that one of the key challenges facing the humanities today is how to transmit the heritage of a culture whose objects are multiplying not simply in mass of items but also in types of system or interface––and where the nature of those varying interfaces greatly complicates the task of identifying, collecting, and otherwise treating the object. This is an enterprise that requires traditional archival research to work in conjunction with Digital Humanities practice where computation methods figure largely (Burdick et al., 3).

Experience

For the last 25 years I have collected works of electronic literature. Driving my efforts besides a fascination with avant-garde literature was the early realization that many of the works in my collection were, over time, becoming impossible to access without computer equipment contemporary with the works themselves. In other words, it became necessary for me to collect, along with floppy disks, CDs, and DVDs holding electronic literary poetry, fiction and essays, computers for which the works were intended to be experienced. To date, I have collected a library of 200 works of electronic literature and 46 vintage computers dating to 1977. I have also found it necessary to collect versions of software for which these works were produced, such as Netscape Communicator used by many electronic literature artists for early experiments with net art and other web-based practices. My library and computer collections are now housed together in the Electronic Literature Lab (ELL, dtc-wsuv.org/wp/ell) at the Vancouver campus. ELL represents the method of digital preservation called Collecting, an approach different from Migrating and Emulating in that it seeks to retain the cultural experience of a work without moving it to a newer platform or representing it in a new setting, respectively.

Bibliography

Electronic Literature Organization.*Electronic Literature Collection*, Vol. **1** and Vol. **2**.http://collection.eliterature.org/

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, **22**(1): 27–47.

Burrows, J. (2004). Textual Analysis. In Schreibman, S., Siemens, R. and Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell.

Hoover, D. (2008). Quantitative Analysis and Literary Studies. In Schreibman, S., Siemens, R. and Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell, pp. 517–33.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Sylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts*. Lincoln (NE): University of Nebraska Lincoln, pp. 487–89.

The Trace of Theory: Extracting Subsets from Large Collections

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca
University of Alberta, Canada

Laura Mandell
mandell@tamu.edu
Texas A&M University

Stéfan Sinclair
stefan.sinclair@mcgill.ca
McGill University

Matthew Wilkens
mwilkens@nd.edu
University of Notre Dame

Boris Capitanu
capitanu@illinois.edu
HathiTrust Research Center

Stephen Downie
jdownie@illinois.edu
University of Illinois, Urbana-Champaign

Introduction

Can we find and track theory, especially literary theory, in very large collections of texts using computers? This panel discusses a pragmatic two-step approach to trying to track and then visually explore theory through its textual traces in large collections like those of the HathiTrust.

1. **Subsetting:** The first problem we will discuss is how to extract thematic subsets of texts from very large collections like those of the HathiTrust. We experimented with two methods for identifying "theoretical" subsets of texts from large collections, using keyword lists and machine learning. The first two panel presentations will look at developing two different types of theoretical keyword lists. The third presentation will discuss a machine learning approach to extracting the same sorts of subsets.

2. **Topic Modelling:** The second problem we tackled was what to do with such subsets, especially since they are likely to still be too large for conventional text analysis tools like Voyant (voyant-tools.org) and users will want to explore the results to understand what they got. The fourth panel presentation will therefore discuss how the HathiTrust Research Center (HTRC) adapted Topic Modelling tools to work on large collections to help exploring subsets. The fifth panel talk will then show an adapted visualization tool, the Galaxy Viewer, that allows one to explore the results of Topic Modelling.

The panel brings together a team of researchers who are part of the "Text Mining the Novel" (TMN) project that is funded by the Social Sciences and Humanities Research Council of Canada (SSHRC) and led by Andrew Piper at McGill University. Text Mining the Novel (novel-tm.ca) is a multi-year and multi-university cross-cultural study looking at the use of quantitative methods in the study of literature, with the HathiTrust Research Center is a project partner.

The issue of how to extract thematic subsets from very large corpora such as the HathiTrust is a problem common to many projects that want to use diachronic collections to study the history of ideas or other phenomena. To conclude the panel, a summary reflective presentation will discuss the support the HTRC offers to DH researchers and how the HTRC notion of "worksets" can help with the challenges posed by creating useful subsets. It will further show how the techniques developed in this project can be used by the HTRC to help other future scholarly investigations.

Using Word Lists to Subset

Geoffrey Rockwell (Kevin Schenk, Zachary Palmer, Robert Budac and Boris Capitanu)

How can one extract subsets from a corpus without appropriate metadata? Extracting subsets is a problem particular to very large corpora like those kept by the HathiTrust (www.hathitrust.org/). Such collections are too large to be manually curated and their metadata is of limited use in many cases. And yet, one needs ways to classify all the texts in a collection in order to extract subsets if one wants to study particular themes, genres or types of works. In our case we wanted to extract theoretical works which for the purpose of this project we defined as Philosophical works or Literary Critical works. In this first panel presentation we will discuss the use of keyword lists as a way of identifying a subset of "philosophical" texts.

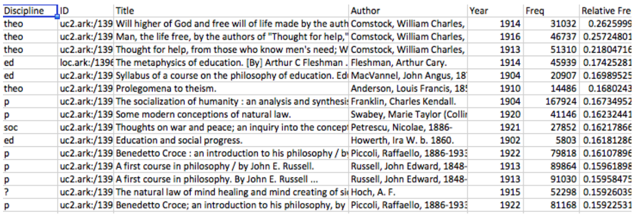
Why philosophical? We choose to experiment extracting philosophical texts first as philosophy is a discipline with a long history and a vocabulary that we hypothesized would lend itself to a keyword approach. Unlike more recent theoretical traditions, philosophical words might allow us to extract works from the HathiTrust going back thousands of years.

Keywords. For the first part of this project we adapted a list of philosophical keywords from the Indiana Philosophy Ontology Project (inpho.cogs.indiana.edu/). Our adapted list has 4,437 words and names starting with "abauzit", "abbagnano", "abdolkarim", "abduction", "abduh", "abel", and so on. There are a number of ways of generating such lists of keywords or features, in our case we were able to start with a very large curated list. The second paper in this panel discusses generating a list of literary critical keywords.

Process. We used this list with a process we wrote in Python that calculates the relative frequency of each word in a text and does this over a collection. The process also calculates the sum of the relative frequencies giving us a simple measurement of the use of philosophical keywords in a text. The word frequency process generates a CSV with the titles, author, frequency sum and individual keyword frequencies which can be checked and manipulated in Excel.

Testing. We iteratively tested this keyword approach on larger and larger collections. First we gathered a collection of 20 philosophical and 20 non-philosophical texts from Project Gutenberg (www.gutenberg.org/). We found the summed frequency accurately distinguished the philosophical from the non-philosophical texts. The process was then run by the HTRC on a larger collection of some 9,000 volumes and the results returned to us. We used the results to refine our list of keywords so that a summed relative frequency of .09 gave us mostly philosophical works with a few false positives. We did this by sorting the false positives by which words contributed to their summed relative frequency and then eliminating those words from the larger list that seemed to be ambiguous.

The process was then run on the HathiTust Open Open collection of 254,000 volumes. This generated some 3230 volumes that had a summed relative frequency over .1, which seemed a safe cut-off point given how .09 had worked with a smaller collection. To assess the accuracy of this method we manually went through these 3,230 and categorized them using the titles producing a CSV that could be used with other classification methods.



Discipline	ID	Title	Author	Year	Freq	Relative Fre
theo	uc2.ark:/139	Will higher of God and free will of life made by the auth	Comstock, William Charles,	1914	31032	0.2625999
theo	uc2.ark:/139	Man, the life free, by the authors of "Thought for help,"	Comstock, William Charles,	1916	46737	0.25724801
theo	uc2.ark:/139	Thought for help, from those who know men's need; W	Comstock, William Charles,	1913	51310	0.21804716
ed	loc.ark:/1396	The metaphysics of education. [By] Arthur C Fleshman .	Fleshman, Arthur Cary.	1914	45939	0.17425281
ed	uc2.ark:/139	Syllabus of a course on the philosophy of education. Edi	MacVannel, John Angus, 18	1904	20907	0.16989525
theo	uc2.ark:/139	Prolegomena to theism.	Anderson, Louis Francis, 18	1910	14486	0.1680243
p	uc2.ark:/139	The socialization of humanity : an analysis and synthesis	Franklin, Charles Kendall.	1904	167924	0.16734952
p	uc2.ark:/139	Some modern conceptions of natural law.	Swabey, Marie Taylor (Collir	1920	41146	0.16232441
soc	uc2.ark:/139	Thoughts on war and peace; an inquiry into the concept	Petrescu, Nicolae, 1886-	1921	27852	0.16217866
ed	uc2.ark:/139	Education and social progress.	Howerth, Ira W. b. 1860.	1902	5803	0.16181286
p	uc2.ark:/139	Benedetto Croce : an introduction to his philosophy / by	Piccoli, Raffaello, 1886-1933	1922	79818	0.16107896
p	uc2.ark:/139	A first course in philosophy / by John E. Russell.	Russell, John Edward, 1848-	1913	89864	0.15961898
p	uc2.ark:/139	A first course in philosophy. By John E. Russell ...	Russell, John Edward, 1848-	1913	91030	0.15958475
?	uc2.ark:/139	The natural law of mind healing and mind creating of sic	Hoch, A. F.	1915	52298	0.15926039
p	uc2.ark:/139	Benedetto Croce; an introduction to his philosophy, by	Piccoli, Raffaello, 1886-1933	1922	81168	0.15922531

The table below summarizes the categories of volumes that we found, though it should be noted that the categorization was based on the titles, which can be misleading. "Unsure" was for works which we weren't sure about. "Not-Philosophical" were those works that we were reasonably sure were not philosophical from the title. The categories like Science and Education were for works about science and philosophy or education and philosophy.

Tag (Type)	Number of Volumes	Example
Unsure	349	The coming revolution (1918)
Education	473	Education and national character (1904)
Philosophy	813	Outlines of metaphysics (1911)
Science	189	Relativity; a new view of the universe (1922)
Social	526	The study of history and sociology (1890)
Religion	722	Prolegomena to theism (1910)
Not Philosophical	158	Pennsylvania archives (1874)

One of the things that stands out is the overlap between religious titles and philosophical ones. This is not surprising given that the fields have been intertwined for centuries and often treat of the same issues. We also note how many educational works and works dealing with society can have a philosophical bent. It was gratifying to find only 4.9% of the volumes classified seemed clearly not philosophical. If one includes the Unsure category it is 15.7%, but the Unsure category is in many ways the most interesting as one reason for classifying by computer is to find unexpected texts that challenge assumptions about what is theory.

Conclusions. Using large keyword lists to classify texts is a conceptually simple method that can be understood and used by humanists. We have lists of words and names at hand in specialized dictionaries and existing classification systems. Lists can be managed to suit different purposes. Our list from InPhO had the advantage that is was large and inclusive, but also the disadvantage that included words like "being" and "affairs" that have philosophical uses but are also used in everyday prose. The same is true of the names gathered like Croce that can refer to the philosopher or the cross (in Italian). Further trimming and then weighting of words/names could improve the classification of strictly philosophical texts. We also need to look deeper into the results to find not just the false positives, but also the true negatives. In sum, this method has the virtue of simplicity and accessibility and in the case of philosophical texts can be used to extract useful, though not complete, subsets.

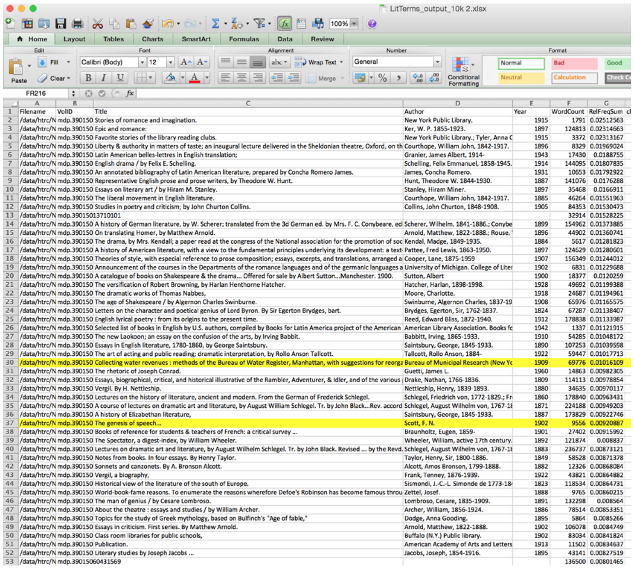
The Problem with Literary Theory

Laura Mandell (Boris Capitanu, Stefan Sinclair, and Susan Brown)

In this short paper, I describe adapting the word list approach developed by Geoffrey Rockwell for extracting a subset of philosophical texts from a large, undifferentiated corpus, to the task of identifying works of literary theory. The degree to which running the list of terms did in fact pull out and gather together works of literary criticism and theory is very high, despite potential problems with such an enterprise, which we discuss in this talk in detail.

1. **Developing the list of literary terms**. Susan Brown and I decided to gather lists of literary terms. Susan initiated a discussion with the MLA about using terms from the *MLA Bibliography* but upon consideration these were in fact not at all what we needed: they classified subjects of texts as opposed to listing terms that would appear in those texts. I had recently spent some time learning about JSTOR's new initiative in which sets of terms are created by what they call "SMEs"--Subject Matter Experts--and then used to locate articles all participating in an interdisciplinary subject. Their first foray is available in Beta: it gathers together all articles in no matter what field on the topic of Environmental Sustainabilty (labs.jstor.org/sustainability/). The terms collected are terms that would appear *in* the relevant texts, not in the metadata about them; the goal is to collect documents across multiple categories related to specialization, discipline, and field, since the desired result to gather together interdisciplinary texts concerning a common topic.

2. **Anachronism.** JSTOR had started a "literary terms" list, and I finished the list of terms relying on encyclopedias of literary theory. Could a list of terms significant in the late-twentieth-century theories of literature as expressed in articles gathered in JSTOR be used to extract a set of texts published much earlier that analyze literature? What about the historical inaccuracy of using twentieth-century terms to find eighteenth- and nineteenth-century literary criticism?



	A	B	C	D	E	F	G	
1	Filename	VolID	Title	Author	Year	WordCount	RelFreqSum	cl
2	/data/htrc/N	mdp.390150	Stories of romance and imagination.	New York Public Library.	1915	1791	0.02512563	
3	/data/htrc/N	mdp.390150	Epic and romance:	Ker, W. P. 1855-1923.	1897	124813	0.02314663	
4	/data/htrc/N	mdp.390150	Favorite stories of the library reading clubs.	New York Public Library.; Tyler, Anna C	1915	3372	0.02313167	
5	/data/htrc/N	mdp.390150	Liberty & authority in matters of taste; an inaugural lecture delivered in the Sheldonian theatre, Oxford, on th	Courthope, William John, 1842-1917.	1896	8329	0.01969024	
6	/data/htrc/N	mdp.390150	Latin American belles-lettres in English translation;	Granier, James Albert, 1914-	1943	17430	0.0188755	
7	/data/htrc/N	mdp.390150	English drama / by Felix E. Schelling.	Schelling, Felix Emmanuel, 1858-1945.	1914	144095	0.01807835	
8	/data/htrc/N	mdp.390150	An annotated bibliography of Latin American literature, prepared by Concha Romero James.	James, Concha Romero.	1931	10653	0.01792922	
9	/data/htrc/N	mdp.390150	Representative English prose and prose writers, by Theodore W. Hunt.	Hunt, Theodore W. 1844-1930.	1887	141076	0.0176288	
10	/data/htrc/N	mdp.390150	Essays on literary art / by Hiram M. Stanley.	Stanley, Hiram Miner.	1897	35468	0.0166911	
11	/data/htrc/N	mdp.390150	The liberal movement in English literature.	Courthope, William John, 1842-1917.	1885	46264	0.01551963	
12	/data/htrc/N	mdp.390150	Studies in poetry and criticism; by John Churton Collins.	Collins, John Churton, 1848-1908.	1905	84353	0.01530473	
13	/data/htrc/N	mdp.39015013710101				32914	0.01528225	
14	/data/htrc/N	mdp.390150	A history of German literature, by W. Scherer; translated from the 3d German ed. by Mrs. F. C. Conybeare, edi	Scherer, Wilhelm, 1841-1886.; Conybe	1899	154962	0.01373885	
15	/data/htrc/N	mdp.390150	On translating Homer, by Matthew Arnold.	Arnold, Matthew, 1822-1888.; Rouse, '	1896	44902	0.01360741	
16	/data/htrc/N	mdp.390150	The drama, by Mrs. Kendall; a paper read at the congress of the National association for the promotion of soc	Kendal, Madge, 1849-1935.	1884	5617	0.01281823	
17	/data/htrc/N	mdp.390150	A history of American literature, with a view to the fundamental principles underlying its development: a text-	Pattee, Fred Lewis, 1863-1950.	1897	124629	0.01280601	
18	/data/htrc/N	mdp.390150	Theories of style, with especial reference to prose composition; essays, excerpts, and translations, arranged a	Cooper, Lane, 1875-1959	1907	156349	0.01244012	
19	/data/htrc/N	mdp.390150	Announcement of the courses in the Departments of the romance languages and of the germanic languages a	University of Michigan. College of Liter	1902	6831	0.01229688	
20	/data/htrc/N	mdp.390150	A catalogue of books on Shakespeare & the drama... Offered for sale by Albert Sutton...Manchester. 1900.	Sutton, Albert	1900	18377	0.0120259	
21	/data/htrc/N	mdp.390150	The versification of Robert Browning, by Harlan Henthorne Hatcher.	Hatcher, Harlan, 1898-1998.	1928	49692	0.01199388	
22	/data/htrc/N	mdp.390150	The dramatic works of Thomas Nabbes,	Moore, Charlotte.	1918	24687	0.01194961	
23	/data/htrc/N	mdp.390150	The age of Shakespeare / by Algernon Charles Swinburne.	Swinburne, Algernon Charles, 1837-19	1908	65976	0.01165575	
24	/data/htrc/N	mdp.390150	Letters on the character and poetical genius of Lord Byron. By Sir Egerton Brydges, bart.	Brydges, Egerton, Sir, 1762-1837.	1824	67287	0.01138407	
25	/data/htrc/N	mdp.390150	English lyrical poetry : from its origins to the present time.	Reed, Edward Bliss, 1872-1940.	1912	178838	0.01133987	
26	/data/htrc/N	mdp.390150	Selected list of books in English by U.S. authors, compiled by Books for Latin America project of the American	American Library Association. Books fo	1942	1337	0.01121915	
27	/data/htrc/N	mdp.390150	The new Laokoon; an essay on the confusion of the arts, by Irving Babbit.	Babbitt, Irving, 1865-1933.	1910	54285	0.01048172	
28	/data/htrc/N	mdp.390150	Essays in English literature, 1780-1860, by George Saintsbury.	Saintsbury, George, 1845-1933.	1890	107253	0.01039598	
29	/data/htrc/N	mdp.390150	The art of acting and public reading; dramatic interpretation, by Rollo Anson Tallcott.	Tallcott, Rollo Anson, 1884-	1922	59447	0.01017713	
30	/data/htrc/N	mdp.390150	Collecting water revenues : methods of the Bureau of Water Register, Manhattan, with suggestions for reorga	Bureau of Municipal Research (New Yc	1909	69776	0.01016109	
31	/data/htrc/N	mdp.390150	The rhetoric of Joseph Conrad.	Guetti, James L.	1960	14863	0.00982305	
32	/data/htrc/N	mdp.390150	Essays, biographical, critical, and historical illustrative of the Rambler, Adventurer, & Idler, and of the various	Drake, Nathan, 1766-1836.	1809	114113	0.00978854	
33	/data/htrc/N	mdp.390150	Vergil. By H. Nettleship.	Nettleship, Henry, 1839-1893.	1880	34635	0.00970117	
34	/data/htrc/N	mdp.390150	Lectures on the history of literature, ancient and modern. From the German of Frederick Schlegel.	Schlegel, Friedrich von, 1772-1829.; Fr	1860	178840	0.00963431	
35	/data/htrc/N	mdp.390150	A course of lectures on dramatic art and literature, by August William Schlegel. Tr. by John Black...Rev. accord	Schlegel, August Wilhelm von, 1767-18	1871	224188	0.00949203	
36	/data/htrc/N	mdp.390150	A history of Elizabethan literature,	Saintsbury, George, 1845-1933.	1887	173829	0.00922746	
37	/data/htrc/N	mdp.390150	The genesis of speech...	Scott, F. N.	1902	9556	0.00920887	
38	/data/htrc/N	mdp.390150	Books of reference for students & teachers of French: a critical survey ...	Braunholtz, Eugen, 1859-	1901	27402	0.00915992	
39	/data/htrc/N	mdp.390150	The Spectator, a digest-index, by William Wheeler.	Wheeler, William, active 17th century.	1892	121874	0.008837	
40	/data/htrc/N	mdp.390150	Lectures on dramatic art and literature, by August Wilhelm Schlegel. Tr. by John Black. Revised ... by the Revd.	Schlegel, August Wilhelm von, 1767-18	1883	236737	0.00873121	
41	/data/htrc/N	mdp.390150	Notes from books. In four essays. By Henry Taylor.	Taylor, Henry, Sir, 1800-1886.	1849	58528	0.00871378	
42	/data/htrc/N	mdp.390150	Sonnets and canzonets. By A. Bronson Alcott.	Alcott, Amos Bronson, 1799-1888.	1882	12326	0.00868084	
43	/data/htrc/N	mdp.390150	Vergil, a biography,	Frank, Tenney, 1876-1939.	1922	43821	0.00864882	
44	/data/htrc/N	mdp.390150	Historical view of the literature of the south of Europe.	Sismondi, J.-C.-L. Simonde de 1773-18	1823	118534	0.00864731	
45	/data/htrc/N	mdp.390150	World-book-fame reasons. To enumerate the reasons wherefore Defoe's Robinson has become famous throu	Zettel, Josef.	1888	9765	0.00860215	
46	/data/htrc/N	mdp.390150	The man of genius / by Cesare Lombroso.	Lombroso, Cesare, 1835-1909.	1891	132298	0.008564	
47	/data/htrc/N	mdp.390150	About the theatre : essays and studies / by William Archer.	Archer, William, 1856-1924.	1886	78514	0.00853351	
48	/data/htrc/N	mdp.390150	Topics for the study of Greek mythology, based on Bulfinch's "Age of fable,"	Dodge, Anna Gooding.	1895	5864	0.0085266	
49	/data/htrc/N	mdp.390150	Essays in criticism. First series. By Matthew Arnold.	Arnold, Matthew, 1822-1888.	1902	106078	0.0084749	
50	/data/htrc/N	mdp.390150	Class room libraries for public schools,	Buffalo (N.Y.) Public library.	1902	83034	0.00841824	
51	/data/htrc/N	mdp.390150	Publication.	American Academy of Arts and Letters	1913	11502	0.00834637	
52	/data/htrc/N	mdp.390150	Literary studies by Joseph Jacobs ...	Jacobs, Joseph, 1854-1916.	1895	43141	0.00827519	
53	/data/htrc/N	mdp.39015060431569				136500	0.00801465	

In fact, results show solidly that this anachronistic list of terms developed by experts do work to gather materials that preceded and fed into, served to develop, the discipline of literary theory. One of two falsely identified texts among the top relevant documents has to do with water distribution systems which had, as part of its most frequent terms, "meter" and "collection," two terms relevant to analyzing the medium and content of poetry. Other false positives are similarly explicable, and, most important, they are rare.

In this paper, we report upon the effects of running these frequent words on very large datasets using both unsupervised to supervised learning.

Machine Learning

Stefan Sinclair (and Matthew Wilkens)

The third panel presentation deals with machine learning techniques to extract subsets. Unsupervised learning techniques allow us to evaluate the relative coherence of theoretical clusters within large textual fields and to identify distinct theoretical subclasses in the absence of any firmly established anatomy of the discipline. For these reasons, we performed unsupervised classification on three corpora: (1) A large collection (c. 250,000 volumes) of mixed fiction and nonfiction published in the nineteenth and twentieth centuries. (2) A subset of that corpus identified by algorithmic and manual methods as highly philosophical. And (3) A subset similarly identified as literary-critical.

In the case of the large corpus, the goal was to identify subsets containing high proportions of philosophy and criticism. For the smaller sets, we sought to produce coherent groupings of texts that would resemble subfields or concentrations within those areas. In each case, we extracted textual features including word frequency distributions, formal and stylistic measures, and basic metadata information, then performed both *k*-means and DBSCAN clustering on the derived Euclidean distances between volumes.

As in past work on literary texts (Wilkens, 105), we found that we were able to identify highly distinct groups of texts, often those dealing with specialized and comparatively codified subdomains, and that we could subdivide larger fields with reasonable but lower accuracy. The model that emerges from this work, however, is one emphasizing continuity over clear distinction. Subfields and areas of intensely shared textual focus do exist, but a systematic view of large corpora in the philosophical and literary critical domains suggests a more fluid conception of knowledge space in the nineteenth and twentieth centuries.

In parallel with the unsupervised classification performed – an attempt to allow distinctive features to emerge without, or with less, bias – we also performed supervised classification, starting with the training set of 40 texts labelled as Philosophical and Other (mentioned in "Using Word Lists to Subset" above). We experimented with several machine learning algorithms and several parameters to determine which ones seemed most suitable for our dataset. Indeed, part of this work was to recognize and and normalize the situation of the budding digital humanist confronting a dizzying array of choices: stoplists, keywords, relative frequencies, TF-IDF values, number of terms to use, Naïve Bayes Multinomial, Linear Support Vector Classification, penalty parameter, iterations, and so on ad infinitum. Some testing is desirable; some guesswork and some craftwork are essential. We reflect on these tensions more in the iPython notebook (Sinclair et al., 2016) and we will discuss them during the presentation as well.

One of the surprises from these initial experiments in machine learning was that using an unbiased list of terms from the full corpus (with stopwords removed) was considerably more effective than attempting to classify using the constrained philosophical vocabulary. Again, this may be because the keywords list was overly greedy.

Just as we experimented with ever-larger corpora for the "Using Lists to Subset" sub-project, the supervised learning subproject broadened its scope gradually in an attempt to identify theoretical texts unknown to us while examining the efficacy of the methodologies along the way. Indeed, the overarching purpose of adopting all three approaches (keyword-based, unsupervised classification, machine learning) was to compare and contrast different ways of studying theory in a large-scale corpus.

Working with HTRC datasets

Boris Capitanu

The fourth panel presentation focuses on working with the HathiTrust and the particular format of HathiTrust texts. Researchers may obtain datasets directly from HathiTrust [1] by making a special request, after having fulfilled appropriate security and licensing requirements. Datasets in HathiTrust and HTRC are available in two different ways:

- via rsync in Pairtree format
- via Data API

According to "Pairtrees for Object Storage (V0.1)" [2], the Pairtree is "a filesystem hierarchy for holding objects that are located within that hierarchy by mapping identifier strings to object directory (or folder) paths, two characters at a time". In the HathiTrust, the objects consist of the individual volume and associated metadata. Volumes are stored as ZIP files containing text files, one text file for each page, where the text file is named by the page number. A volume ZIP file may contain additional non-page text files, whose purpose can be identified from the file name. The metadata for the volume is encoded in METS XML [3] and lives in a file next to the volume ZIP file. For example, a volume with id "loc.ark:/13960/t8pc38p4b" is stored in Pairtree as:

loc/pairtree_root/ar/k+/=1/39/60/=t/8p/c3/8p/4b/ark+=13960=t8pc38p4b/ark+=13960=t8pc38p4b.zip
loc/pairtree_root/ar/k+/=1/39/60/=t/8p/c3/8p/4b/ark+=13960=t8pc38p4b/ark+=13960=t8pc38p4b.mets.xml

where "loc" represents the 3-letter code of the library of origin (in this case Library of Congress). As mentioned, the volume ZIP files contain text files named for the page number. For example, here are the first few entries when listing the contents of the above ZIP file:

ark+=13960=t8pc38p4b/
ark+=13960=t8pc38p4b/00000001.txt
ark+=13960=t8pc38p4b/00000002.txt
ark+=13960=t8pc38p4b/00000003.txt
…

Note that the strings that encode the volume id and the ZIP filename are different. Before a volume id can be encoded as a file name, it goes through a "cleaning" process that converts any character that is not a valid character to be used in a filename into one that is (for example ":" was converted to "+" and "/" to "="), also dropping the 3-letter library code. The specific conversion rules are obscure, but library code already exists [4][5] for multiple languages that is able to perform this conversion both ways.

The pairtree is an efficient structure for storing a large number of files. However, working with this structure can pose certain challenges. One of the issues is that this deeply nested folder hierarchy is slow to traverse. Applications needing to recursively process the volumes in a particular dataset stored in pairtree will have to traverse a large number of folders to "discover" every volume. A second inconvenience stems from the use of ZIP to store the content of a volume. While efficient in terms of disk space usage, it's inconvenient when applications need to process the text data of the volume as they would need to uncompress the ZIP file and read its contents, in the proper order, concatenating all pages, in order to obtain the entire volume text content. A further complication is due to the fact that the exact ordering and naming of the page text files in the ZIP file is only provided as part of the METS XML metadata file. So, if the goal is to create a large blob of text containing all the pages of a volume (and only the pages, in the proper order, without any additional non-page data), the most correct way of doing so is to first parse the METS XML to determine the page sequence and file names, and then uncompress the ZIP file concatenating the pages in the exact sequence specified. This, of course, has a large performance penalty if it needs to be done on a large dataset every time this dataset is used to address some research question.

An alternative way to obtain a particular dataset is to use the Data API [6]. Currently, access to Data API is limited, and is allowed only from the Data Capsule [7] while in Secure Mode. Using the Data API a researcher can retrieve multiple volumes, pages of volumes, token counts, and METS metadata documents. Authentication via the OAuth protocol is required when making requests to the Data API. The advantage of using the Data API in place of the pairtree (other than disk storage savings) is that one can request already-concatenated text blobs for volumes, and make more granular requests for token counts or page ranges without having to traverse deeply-nested folder structures or parse METS metadata.

In this panel presentation we will show how the tools developed for the Trace of Theory project were adapted to work with the Pairtree format. The goal is to help others be able to work with the HathiTrust data format.

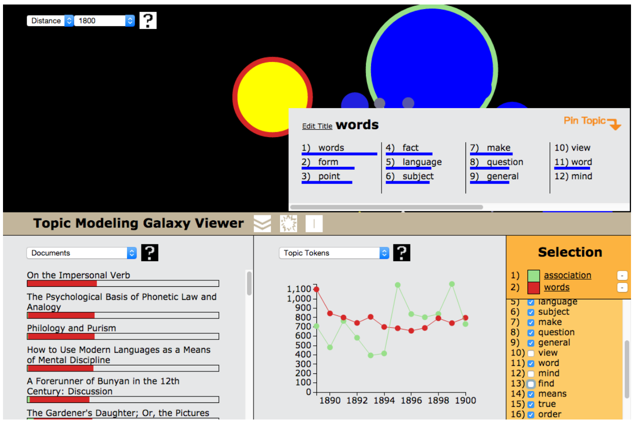
Notes

[1] https://www.hathitrust.org/datasets
[2] http://tools.ietf.org/html/draft-kunze-pairtree-01
[3] http://www.loc.gov/standards/mets/
[4] https://confluence.ucop.edu/display/Curation/PairTree
[5] https://github.com/htrc/HTRC-Tools-PairtreeHelper
[6] https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+API+Users+Guide
[7] https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule

Topic Modelling and Visualization for Exploration

Susan Brown (Geoffrey Rockwell, Boris Capitanu, Ryan Chartier, and John Montague)

When working with very large collections even subsets can be too large to manage with conventional text analysis tools. Further, one needs ways of exploring the results of extraction techniques to figure out if you got what you were expecting or something surprising in an interesting way. In the fifth panel presentation we will discuss the adaptation of a tool called the Galaxy Viewer for visualizing the results of Topic Modelling (Montague et al., 2015). Topic modeling is an automated text mining technique that has proven popular in the humanities that tries to identify groups of words with a tendency to occur together within the same documents in a corpus. Chaney and Blei explain that, “One of the main applications of topic models is for exploratory data analysis, that is, to help browse, understand, and summarize otherwise unstructured collections.” (Chaney et al., 2012)



On the Impersonal Verb

The Psychological Basis of Phonetic Law and Analogy

Philology and Purism

How to Use Modern Languages as a Means of Mental Discipline

A Forerunner of Bunyan in the 12th Century: Discussion

The Gardener's Daughter; Or, the Pictures

The Galaxy Viewer prototype was developed to explore the results of topic modelling over large collections. It combines different views so that one can select topics, compare topics, explore the words in topics, follow topic tokens over time, and see the document titles associated with topics. In this presentation we will demonstrate the Galaxy Viewer and then discuss how it was scaled to handle much larger collections.

The prototype Galaxy Viewer backend code uses Mallet (McCallum, 2002) to infer the set of topics, topic distributions per document, and word probabilities per topic. Unfortunately, Mallet is meant to be used on small- to medium-sized corpora as it requires that the entire dataset be loaded into RAM during training. An additional constraint with Mallet is the fact that although Mallet can fully utilize all the CPU cores on a single machine, it’s not designed to work in a distributed-computing fashion across a number of machines, to speed up execution. As such, processing very large datasets (if even possible) might take a very long time (as the algorithm makes multiple passes over the entire dataset). Many implementations of LDA exist, which primarily fall into one of two categories: Batch LDA, or Online LDA. The core difference between batch and online LDA stems from what happens during each iteration of the algorithm. In batch mode, as mentioned earlier, each iteration of the algorithm makes a full pass over all the documents in the dataset in order to re-estimate the parameters, checking each time for convergence. In contrast, online LDA only makes a single sweep over the dataset, analyzing a subset of the documents each iteration. The memory requirement for online LDA depends on the chosen batch size only, not on the size of the dataset - as is the case with batch LDA.

We are currently in the process of researching/comparing the available implementations of LDA to establish which one would be best suited to use for the Galaxy Viewer. We are also considering the option of not fixing the LDA implementation, but instead make the backend flexible so that any LDA implementation can be used (as long as it provides the appropriate results that are needed). In the latter case we’d have to create specific result interpreters that can translate the output from the specific implementation of LDA to the appropriate format to be used to store in the database (to be served by the web service).

Given that Topic Modeling results do not expose the textual content of the documents analyzed, and cannot be used to reconstruct the original text, they are safe to be publicly shared without fear of violating copyright law. This is great news for researchers working with collections like those of the HathiTrust as they should be able to gain insight into datasets which are still currently in-copyright and would, otherwise, not be available to be inspected freely.

In the prototype Galaxy Viewer implementation, the output of the topic modeling step is processed through a set of R functions that reshape the data and augment it with additional calculated metrics that are used by the web frontend to construct the visualization. These post-processing results are saved to the filesystem as a set of five CSV files. One of these CSV files is quite large as it contains the topic modeling state data from Mallet (containing topic assignments for each document and word, and associated frequency count). The visual web frontend code loads this set of five files into memory when the interface is accessed the first time, which can take several minutes. For the prototype this approach was tolerated, but it has serious scalability and performance issue that needs to be addressed before the tool can be truly usable by other researchers.

Scaling the Galaxy Viewer therefore consists of creating a web service backed with a (NoSQL) database which will service AJAX requests from the front-end for the data needed to construct the topic visualization and related graphs. We are developing the set of service calls that need to be implemented/exposed by the web service to fulfill the needs of the front-end web-app. The backend service will query the database to retrieve the necessary data to service the requests. The database will be created based on the output of the Topic Modeling process, after required post-processing of the results is completed (to calculate the topic trends, topic distances, and other metrics used in the display). Relevant metadata at the volume and dataset level will also be stored to be made available to the front-end upon request. This work will be completed by the end of December 2015 so that it can be demonstrated in the new year. The scaled Galaxy Viewer will then provide a non-consumptive way of allowing users of the HathiTrust to explore the copyrighted collections. Extraction of subsets and Topic Modelling can take place under the supervision of the HTRC and the results database can then be exposed to visualization tools like the Galaxy Viewer (and others) for exploration.

Closing reflections: How "Trace of Theory" will improve the HTRC

J. Stephen Downie

The HathiTrust Research Center exists to give the Digital Humanities community analytic access to the HathiTrust's 13.7 million volumes. The HT volumes comprise over 4.8 billion pages each in turn represented by a high-resolution image file and two OCR files yielding some 14.4 billion data files! Thus, as the earlier papers have highlighted, the sheer size of the collection, along with the idiosyncratic nature of the HT data, together create several hurdles that impede meaningful analytic research. The HTRC is engaged in two ongoing endeavours designed to assist DH researchers in overcoming these obstacles: The Advance Collaborative Support (ACS) program [1]; and, the Workset Creation for Scholarly Analysis (WCSA) project [2].

The ACS program at HTRC provides no-cost senior developer time, data wrangling assistance, computation time and analytic consultations to DH researchers who are prototyping new research ideas using the HT data resources. The ACS program is an integral part of the HTRC's operation mission and was part of its value-added proposition when the HTRC launched its recent four-year operations plan (2014-2018). It is a fundamental component of the HTRC's outreach activities and as such, has staff dedicated to its planning, management and day-to-delivery. The ACS team was responsible for creating, and then reviewing, the competitive ACS Request for Proposals (RFP) that ask interested DH researchers outline their intellectual goals, describe their data needs, and estimate their computational requirements. The ACS team is generally looking for new projects that could benefit from some kickstarting help from HTRC. HTRC welcomes proposals from researchers with a wide range of experience and skills. Projects run 6 to 12 months.

Originally funded by the Andrew W. Mellon Foundation (2013-2015), the current WCSA program is building upon, extending and implementing the development made during the funding period. The HTRC project team, along with subaward collaborators at University of Oxford, University of Maryland, Texas Agriculture and Marine University and University of Waikato, developed a group of prototype techniques for empowering scholars who want to do computational analyses of the HT materials to more efficiently and effectively create user-specific analytic subsets (called "worksets"). A formal model has been designed to describe the items in a workset along with necessary bibliographic and provenance metadata that is now being incorporated into the HTRC infrastructure (Jett, 2015).

The Trace of Theory project was selected from the first round of ACS proposals. This concluding panel presentation will discuss in what ways the Trace of Theory project has been both a representative and a unique exemplar of the ACS program. It will present some emergent themes that evolved from the HTRC-Trace of Theory interactions that we believe will have an important influence on the delivery of future ACS projects. In the same manner, it will reflect upon the problems the team of researchers had in subsetting the data to build their necessary worksets along with the solutions that the HRTC-Trace of Theory collaboration developed to surmount those difficulties. The panel will finish with a summary of how HTRC intends to incorporate the lessons learned into its day-to-day operations as well as future ACS projects.

Notes:

[1] The 2014 ACS RFP is available at: https://www.hathitrust.org/htrc/acs-rfp

[2] https://www.lis.illinois.edu/research/projects/workset-creation-scholarly-analysis-prototyping-project

Bibliography

Chaney, A. J. and Blei, D. M. (2012). *Visualizing Topic Models*, ICWSM. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4645%26lt%3B/5021 (accessed Dec 2015).

Jett, J. (2015). *Modeling worksets in the HathiTrust Research Center*. CIRSS Technical Report WCSA0715. Champaign, IL: University of Illinois at Urbana-Champaign. Available via: http://hdl.handle.net/2142/78149 (accessed Dec 2015)

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*, http://mallet.cs.umass.edu (accessed Dec 2015).

Montague, J., Simpson, J., Brown, S., Rockwell, G. and Ruecker, S. (2015). Exploring Large Datasets with Topic Model Visualization. *Paper presented by Montague at DH 2015 at the University of Western Sydney*, Australia.

Sinclair, S., G. Rockwell and the Trace of Theory Team.

(2016). **Classifying Philosophical Texts**. Online at http://bit.ly/1kHBy56 (accessed Dec 2015).

Wilkens, M. (2016). Genre, Computation, and the Weird Canonicity of Recently Dead White Men. *NovelTM Working Paper*.

Digital Folkloristics: the Use of Computational Methods in Revealing the Characteristics of Folkloric Communication

Mari Sarv
mari@haldjas.folklore.ee
Estonian Literary Museum, Estonia

Liisi Laineste
digitaalhumanitaaria@gmail.com
Estonian Literary Museum, Estonia

Greta Franzini
gfranzini@gcdh.de
Göttingen Centre of Digital Humanities, Germany

Emily Franzini
efranzini@gcdh.de
Göttingen Centre of Digital Humanities, Germany

Kati Kallio
kati.kallio@gmail.com
Finnish Literary Society

Risto Järv
risto@folklore.ee
Estonian Literary Museum, Estonia

Panel Topic

Folkloric communication is a specific mode of social interaction that uses existing knowledge and creatively adapts it to particular situations, audience, and intent. It always includes some repetition and some innovation or, in other words, as Michel de Certeau (1984) expresses it in his study of everyday practices, the main attribute of cultural transmission is the changing nature of everything that is being passed on.

The questions of authorship, stylistics and variation in time and space have been discussed and problematized in folklore studies for over a hundred years. Digital era has brought along mass digitization of cultural heritage documents and the compilation of folklore databases and text corpora (see e.g. Schmitt, 2014). However, the use of computational methods in researching folklore and its specifics has thus far been modest. The intention of this panel is to discuss the possibilities that computational methods have to offer in revealing the inherent qualities of folkloric communication in various text collections and research corpora. The participants of the panel deal with different source data including archival records and publications as well as contemporary social media. They discuss different aspects of folkloric communication and variation, contributing together to a general insight into these processes.

The Panel: speakers and contributions

In this panel, the participants will introduce their projects, data, methods and research results to contextualize and elicit a discussion on the state-of-the-art in digital folkloristics, and on possible ways forward.

Mari Sarv and **Risto Järv** (Estonian Literary Museum) analyse the essence of **folkloric variation** relying on the text corpora from the collections of Estonian Folklore Archives. The Estonian Runic Songs' database (1996-2015) contains ca. 100,000 poetic texts, and the database of folk tales consists of 11,000 tales (both together with metadata). Previous studies of folksongs have shown that the statistical analysis of poetical features of songs as well as their content allows us to locate their geographical origin (i.e. tracing the belonging of songs to the tradition of a local community) quite precisely. At the same time there are clear differences in the geographical distribution of linguistic-poetical features, especially compared to the elements of content. The stylistic analysis of folklore texts enables us to find out how much the personal style of performer is revealed in folkloric recordings of traditional plots ('types' in the folkloristic discourse). The potential of computational methods to tackle the dichotomy of stability and variation in the folkloric communication poses a most intriguing challenge.

Kati Kallio (Finnish Literary Society) will discuss **the characteristics of oral poetry**, including complex patterns of variation and tricky definitions of authorship, and concentrate on the challenges posed by of the SKVR-corpus of Finnic oral poetry (see also Digital Archive of Finnish Folk Tunes). The corpus represents various languages and dialects, orthographies, personal writing styles and traces of different modes of performance, but is held together by similar poetic registers. For this kind of specific poetic register in several related small languages and across a wide variety of genres, no ready-made tools for computational linguistic analysis exists. On the other hand, the corpus already includes a detailed thematic index, and the researchers have applied various manual methods to the corpus for hundred years. What kinds of new questions could be answered with computational methods? The author will present some test analyses and discuss the future possibilities of digital folkloristics on oral poetry.

Greta Franzini and **Emily Franzini** (Göttingen Centre for Digital Humanities) will elaborate on two research projects focusing on the computational interrogation of folktale collections and corpora. Using the Brothers Grimm's Kinder- und Hausmärchen as a case study and base reference, one project addresses the popularization of fairy tale motifs triggered by the Brothers, and seeks to algorithmically crawl web corpora to study the global network of motifs. Motifs form the significant set of "key words" of a tale. In order to systematically crawl for parallel texts, we prepare a digital, machine-readable and -citable index in different languages. TRACER as a text reuse framework is used to check if a document contains similar keywords or describes the same tale. It implements a seven-layer approach combining segmentation, preprocessing, featuring, selection, linking, scoring, and postprocessing steps. The other project examines the textual evolution of the Kinder- und Hausmärchen, starting with the first edition published in 1812 to the seventh and last in 1857. The number of fairy tales grew with every edition, and the numerous changes the Brothers made over the decades in terms of both style and content were symptomatic of societal interest and development. These seven editions represent an ideal testbed not only to computationally verify existing research about this progression but also to identify and distinguish the authorial and stylistic fingerprints of Jacob and Wilhelm Grimm. The discussion will demonstrate the possibilities afforded by the Digital Humanities to conduct **web-scale** and **Big Data** research.

Liisi Laineste (Estonian Literary Museum) is applying the methodologies of Digital Humanities to the folkloric aspect of **social media** content in combination with theories of global information flow, participatory journalism and humour theory. Internet has become central in contemporary cultural communication – most of online communication can be treated as folklore, in which shared norms and values are constructed through cultural artifacts. Forums, commentary boards of news sites, blogs, Twitter and other social media applications have become commonplace during the last fifteen years. In the Internet, news, ideas and opinions travel fast. Social media spreads folklore in unforeseen volumes across national and cultural boundaries. Above all, the focus is on producing and consuming texts, images and multimedia. The presentation will attempt to trace such cultural texts as they move across and between cultures. Besides, as social media is often perceived as a catalyst and accelerator of public discussion and citizen movements, cultural texts as valuable agents in citizen engagement will be discussed in the light of the 2015 refugee crisis in Europe.

Bibliography

De Certeau, M. (1984). *The Practice of Everyday Life*. Berkeley: University of California Press.

Digital Archive of Finnish Folk Tunes, http://esavelmat.jyu.fi/index_en.html (accessed 27 October 2015).

Estonian Runic Songs' database (1996-2015), Estonian Folklore Archives. Estonian Literary Museum, 1996-2015, http://www.folklore.ee/regilaul/andmebaas/?ln=en (accessed 1 November 2015).

Schmitt, Christoph (Ed.) (2014). *Corpora ethnographica online. Strategies to digitize ethnographical collections and their presentation on the Internet*. Waxmann Verlag GmbH. (Rostocker Studien zur Volkskunde und Kulturgeschichte; 5).

SKVR-tietokanta – kalevalaisten runojen verkkopalvelu. Suomalaisen Kirjallisuuden Seura, http://skvr.fi. (accessed 27 October 2015).

APIs in Digital Humanities: The Infrastructural Turn

Toma Tasovac
ttasovac@humanistika.org
Belgrade Center for Digtial Humanities

Adrien Barbaresi
adrien.barbaresi@oeaw.ac.at
Austrian Academy of Sciences

Thibault Clérice
thibault.clerice@uni-leipzig.de
University of Leipzig

Jennifer Edmond
jedmond36@gmail.com
Trinity College Dublin

Natalia Ermolaev
nataliae@Princeton.EDU
Princeton University

Vicky Garnett
garnetv@tcd.ie
Trinity College Dublin

Clifford Wulfman
cwulfman@Princeton.EDU
Princeton University

As a community of practice, digital humanists deal with data and metadata not as static artifacts, but rather as complex, multi-dimensional and multi-layered datasets that can be analyzed, annotated and manipulated in order to produce new knowledge. One of the most important challenges facing DH today is how to consolidate and repurpose available tools; how to create reusable but flexible workflows; and, ultimately, how to integrate and disseminate knowledge, instead of merely capturing it and encapsulating it. This technical and intellectual shift can be seen as the "infrastructural turn" in digital humanities (Tasovac et al. 2015).

Application Programming Interfaces (APIs) have the potential to be powerful, practical building blocks of digital humanities infrastructures. On the technical level, they let heterogeneous agents dynamically access and reuse the same sets of data and standardized workflows. On the social level, they help overcome the problem of "shy data", i.e. data you can "meet in public places but you can't take home with you" (Cooper 2010). Some 10 years ago, Dan Cohen started the conversation about APIs in DH by pointing out that, despite their potential, Andreas few humanities projects — in contrast to those in the sciences and commercial realms — were developing APIs for their resources and tools (Cohen 2005). In the decade since, API development in the digital humanities has certainly increased: today, both large-scale, national and international initiatives, such as HathiTrust, DPLA or Europeana, as well as individual projects, such as Canonical Text Services (CTS), Open Siddur, Folger Digital Texts, correspSearch etc., are focusing their attention and resources on developing APIs. It is now time to reflect on this development: have standards or best-practices evolved? What workflows are most effective and efficient for creating APIs? What are the challenges or stumbling blocks for creating or using APIs? Are APIs being used by DH researchers? What is the future of API development and use in the humanities community?

This panel will cover both the theory and practice of APIs in the digital humanities today. It will bring together researchers working on major European and North American projects, who will discuss APIs from the perspectives of design, implementation, and use, as well as technical and social challenges. Each group will have 10 minutes for their statement, and 40 minutes will remain for group discussion and questions from the audience. One of the panel members will serve as the moderator. All speakers have confirmed their intention to participate in the panel.

Toma Tasovac (Belgrade Centre for Digital Humanities) will discuss an API-centric approach to designing and implementing digital editions. Starting with the notion of text-as-service and textual resources as dynamic components in a virtual knowledge space, Tasovac will show how two recent projects — *Raskovnik: A Serbian Dictionary Platform* and *Izdanak: A Platform for Digital Editions of Serbian Texts* — were implemented using API-focused data modeling at the core of the project design process. The API-first approach to creating TEI-encoded digital editions offers tangible interfaces to textual data that can be used in tailor-made workflows by humanities researchers and other users, well-suited to distant reading techniques, statistical analysis and computer-assisted semantic annotation. The "infrastructural turn" in Digital Humanities does not only have practical implications for the way we build tools and create resources, but also has theoretical ramifications for the way we distinguish highly from loosely structured data: if text is not an object, but a service; and not a static entity, but an interactive method with clearly and uniquely addressable components, a formal distinction between a dictionary and, say, a novel or a poem, is more difficult to maintain.

Clifford Wulfman and **Natalia Ermolaev** (Center for Digital Humanities, Princeton) will discuss the design and implementation of Blue Mountain Springs, the API for the Blue Mountain Project's collection of historic avant-garde periodicals. By modeling magazine data using the FRBRoo ontology and its periodical-oriented extension PRESSoo (PRESSoo, 2014), this RESTful API exposes the Blue Mountain resource in a variety of data formats (structured metadata, full-text, image, linked data). The authors will provide several examples of how Blue Mountain Springs has been used by researchers, drawing especially from the results of the hackathon they will host at Princeton in February 2016, which will bring together approximately twenty periodical studies scholars, technologists, and librarians to work with the API. Creating APIs is part of a trend in DH to move into a post-digital-library phase, when the traditional library functions of discovery and access are no longer sufficient to support research in the humanities. This trend also suggests that DH researchers must reconceptualize their own engagement with material, to think less in terms of monographs and more in terms of resources, and consequently to promulgate their work not as web sites but as web services.

Thibault Clérice (University of Leipzig) will discuss the design of the Canonical Text Services (CTS) and its URN scheme, which make the traditional citation system used by classicists machine-actionable (Blackwell and Smith 2014)[1]. The Homer Multitext (HMT) implementation of CTS requires textual data to be extracted out of its original digital representation into RDF triples in order to be served. The Perseus Digital Library (PDL) implementation, on the other hand, uses extended transformations to slice XML files into multiple records, each representing a passage at a certain level. While relational and RDF database approaches have had some success in scalability and speed (Tiepmar 2015), they also have to deal with maintenance and evolution capacity. There is a real need for this type of DH projects to scale not only in terms of data retrieval speeds, but also in terms of allowing researchers to correct and enhance their data. In addition, projects need to be able to propose other narratives: sliced data doesn't easily provide access to the full data model. Clérice will discuss why and how, using both a native XML-based system such as eXist and a Python-based implementation, one can achieve scalability while guaranteeing maintenance and evolution.

Adrien Barbaresi from the Austrian Academy of Sciences (ICLTT) will discuss the use of APIs in building resources for linguistic studies. The first case deals with lesser-known social networks (Barbaresi 2013) while the second tackles the role of the Twitter API in building the ICLTT's "tweets made in Austria" corpus[2]. For computational linguists, short messages published on social networks constitute a "frontier" area due to their dissimilarity with existing corpora (Lui & Baldwin 2014), most notably with reference corpora of written language. Since data are mainly accessed and collected through APIs and not in the form of web pages, Barbaresi argues that social networks are a frontier area for (web) corpus construction. He will point out the challenges of using Twitter's API, for example how to reveal the implicit decisions and methodology used by API designers, as well as concrete implementation issues, such as the assessment and optimization of data returned by the API. Free APIs may come at no cost, but they also offer no guarantee, so that the use of commercial APIs for research purposes has to be seen with a critical eye in order to turn a data collection process into a proper corpus.

Finally, **Jennifer Edmond** and **Vicky Garnett** (Trinity College Dublin), will provide reflections on the place of APIs within European research infrastructures for the humanities. Their contribution to the panel builds on their recent study on the Europeana Cloud project, which found that while access to data is a real and growing area of interest, very few humanities researchers seem to actively and directly use APIs.[3] They will describe two initiatives, one technical, one social, aiming to better harness the potential of the API to meet researcher's implicit needs. The first is the Collaborative European Digital Archival Research Infrastructure (CENDARI) project, whose platform is structured around an internal API that will allow multiple data sources (local repository, triple store, metasearch engine) to be aligned, enhanced and then served out to a number of environments and tools, including the project's native note-taking environment. The second example is the genesis and development of the concept of the 'inside-out' archive. This framework, which has arisen out of a collaborative venture between several European humanities research infrastructure projects, seeks to encourage collection holding institutions to look beyond their own digitization programs and platforms and recognize the rising importance of machines-as-users (requiring specific access points and formats) rather than the somewhat outdated model of individual institutional web presence serving individual human resource seekers.

The five speakers on this panel will address some of the most pressing issues related to the ongoing development and future of APIs on the DH research infrastructure landscape. The discussion will cover both micro- and macro levels, ranging from methodological implications and technical scalability to the ways in which API-based data access to collections challenges traditional norms of institutional identity and independence. As such, the panel will offer a timely platform for a multifaceted debate on the potentials and pitfalls of building and using APIs in the digital humanities.

Bibliography

Badenoch, A. and Fickers A. (2010). Europe Materializing? Toward a Transnational History of European Infrastructures. In Badenoch, A. and Fickers A. (eds.), Materializing Europe: Transnational Infrastructures and the Project of Europe, 1-26. Basingstoke, Hampshire; New York: Palgrave Macmillan.

Barbaresi, A. (2013). Crawling microblogging services to gather language-classified URLs. Workflow and case study. In Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop, pages 9-15.

Blackwell, C. and Smith N. (2014). *Canonical Text Services Protocol Specification*. http://folio.furman.edu/projects/citedocs/cts/. Accessed October 23, 2015.

Cohen, D. Do APIs Have a Place in the Digital Humanities. http://www.dancohen.org/blog/posts/do_apis_have_a_place_in_the_digital_humanities. Accessed October 24, 2015.

Cohen, D. (2006). "From Babel to Knowledge: Data Mining Large Digital Collections." *D-Lib Magazine* 12, no. 3.

Cooper, D. (2010). When Nice People Won't Share: Shy Data, Web APIs, and Beyond, *Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, n. pag.

Edmond, J, Bulatovic N. and O'Connor, A. "The Taste of 'Data Soup' and the Creation of a Pipeline for Transnational Historical Research." Journal of the Japanese Association for Digital Humanities 1, no. 1 (2015): 107–22.

Edmond, J. and Garnett V. (2015). APIs and Researchers: The Emperor's New Clothes, International Journal of Digital Curation, 10(1): 287-97.

LeBeuf, Patrick (ed.). PRESSoo. Extension of CIDOC CRM and FRBROO for the modelling of bibliographic information pertaining to continuing resources. Version 0.5, http://www.ifla.org/files/assets/cataloguing/frbr/pressoo_v0.5.pdf. Accessed, November 1, 2015.

Murdock, J. and Allen C. (2011). InPhO for All: Why APIs Matter, Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1(3): http://www.jamram.net/docs/jdhcs11-paper.pdf. Accessed, October 23, 2015.

Tasovac, T. Rudan S. and Rudan S. (2015). Developing Morpho-SLaWS: An API for the Morphosyntactic Annotation of the Serbian Language. Systems and Frameworks for Computational Morphology, 137-47. Heidelberg: Springer.

Tiepmar, J. (2015). Release of the MySQL based implementation of the CTS protocol. In Bański, P., Biber H., Breiteneder E., Kupietz M., Lüngen H. and Witt A. (eds.), Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), 35-43. Mannheim: Institut für Deutsche Sprache.

Notes

[1] Infrastructures are installations and services that function as "mediating interfaces" or "structures 'in between' that allow things, people and signs to travel across space by means of more or less standardized paths and protocols for conversion or translation" (Badenoch and Fickers 2010, 11). Digital research infrastructures are no different: they are a mediating set of technologies for research and resource discovery, collaboration, sharing and dissemination of scholarly output.

[2] http://bluemountain.princeton.edu

[3] Traditional scholars have been citing texts the same way for centuries : for example, "Hom. Il. 1.1", which corresponds to Homer's Iliad, Book 1 Line 1. However, although the passage identifier does not change from one scholar to another in most cases, the abbreviation used for the author and the work title will diverge among authors, countries, and publications.

[4] http://www.oeaw.ac.at/icltt/node/193

[5] This is a result of the confluence of several factors: that the research data most humanists want to access is seldom available via an API; that many sources that do offer relevant data lack the structure or metadata to make the API useful; and that humanistic researcher generally lack the skill set to experiment directly with APIs. At the same time, however, one of the key desires expressed by researchers in the current landscape is the federation of high-quality data (Edmond and Garnett 2015).

MEDEA (Modeling Semantically Enriched Digital Editions of Accounts)

Kathryn Tomasek
tomasek_kathryn@wheatoncollege.edu
Wheaton College, Norton, Massachusetts, United States of America

Georg Vogeler
georg.vogeler@uni-graz.at
Centre for Information Modeling - Austrian Centre for Digital Humanities at Karl University, Graz, Austria

Kathrin Pindl
Kathrin.Pindl@geschichte.uni-regensburg.de
University of Regensburg, Germany

Clifford Anderson
clifford.anderson@vanderbilt.edu
Vanderbilt University, United States

Anna Paulina Orlowska
anna.p.orlowska@gmail.com
University of Kiel, Germany

Øyvind Eide
oe@oeide.no
University of Passau, Germany

Most human communities have produced accounts of various sorts, and scholars have long used accounts as primary sources for economic and social history. MEDEA is a collaborative project that involves participants from three continents and includes projects using primary sources that span several centuries across multiple geographical regions. We recommend that digital editions of accounts—business, governmental, personal—use XML/TEI, the widely accepted stable archival format for digital scholarly editions, as a first step. XML/TEI-based editions of accounts should be further encoded using RDFs/OWL to take advantage of the affordances of the Semantic Web and Linked Open Data. Producing semantically enhanced digital scholarly editions of accounts on this general model will create opportunities to ask new questions about the social and economic lives of people in the past. The session features participants from MEDEA's first year; we expect the project to continue for several more years, depending on resources.

Accounts: abstract models of human interactions

Account books share with printed and manuscript texts in prose, poetic, or dramatic forms a problem in data modeling because they are already abstract representations of human interactions. In the case of accounts, although these representations have taken different forms in different times and places, they have an apparent uniformity in that they often consist of lists of people, goods, and services accompanied by numerical values that represent amounts of currency, credit, or in-kind benefits exchanged. Often but not always, these values are totaled or balanced in some way (Tomasek and Bauman, 2013). The development of printed ledgers and spreadsheets illustrates the apparent uniformity of the structure of account books. Contemporary spreadsheet software fails to recognize the conceptual abstractions embedded in the physical format, but these unmarked abstractions are carried into the representations of spreadsheet information in digital form.

Accounts contain both numerical and semantic information (Thaller, 2012; Vogeler, 2014). Various types of accounts lend insight into business dealings, institutional, municipal, and state activities as well as personal uses of goods, property, and currency. The numerical information might include prices, wages, payments of customs duties or taxes as well as quantities of time worked and goods exchanged. In its simplest form, such information can be input as comma separated values (csv), and spreadsheet software can be used to perform calculations on the data. Through the emphasis on calculation, spreadsheets privilege the numerical information found in accounts.

But financial records include information of historical interest in addition to data expressed in numerical form. Account books often include the names of people with whom an individual or organization exchanged goods and services. Goods, services, and currencies have historically significant values in addition to the numerical data expressed as prices, wages, and quantities of time or goods, and such values have associations that are of historical interest. For example, different grains were staples for bread making in different regions (Allen, 2001).

Indeed, the semantic information contained in accounts can provide a rich picture of a community at the local level as well as that community's relationships to others across space and time. In combination with other kinds of community records well known to social and economic historians—tax lists, church memberships, probate inventories, manuscript census returns—the semantic information embedded in accounts can reveal insights that the pure numerical representation might obscure. And such information might open up new avenues of inquiry or facilitate new insights into the operations of past networks of human relationships.

Historical accounts on the Semantic Web

If the semantic information is marked in standardized machine-readable ways, digital editions of accounts can be combined with Semantic Web technologies to produce much richer comparative information about the human relationships they represent than has been possible with earlier iterations of social and economic history. In fact, accounts are an example of a problem space in which insights of

Digital Humanities can improve longstanding practices in social scientific uses of computers to produce data about past human practices and relationships.

Since the Semantic Web offers opportunities to collect and compare data from multiple digital projects, the MEDEA project looks to the potential of developing broad standards for producing semantically enriched digital editions of accounts. Because the Guidelines of the Text Encoding Initiative (TEI) offer a method for creating stable humanities-oriented data from textual sources, MEDEA explores models for building on them to test ways to publish data on commodities, wages, and prices susceptible to comparative analysis.

While the TEI Guidelines provide a standard for markup of manuscript and print sources, some of the elements and attributes most useful for accounts fail to model machine-readable values adequate to the goal of comparability of accounts originally created across broad ranges of time and space. Thus the MEDEA project looks also to CIDOC-CRM and RDF/OWL as sites to begin consideration of the kinds of taxonomies and ontologies that will produce standard machine-readable values to express some of the semantic values found in accounts--especially information about commodities and currencies--that are relevant to humanities scholars. This includes conversion of local measures (Vogeler, 2014).

At present, the MEDEA leadership team imagines semantically enriched digital editions as networks of references between several digital representations of original archival account books. Such editions will allow scholars with varied interests to use the data from the accounts in different ways according to their fields and scholarly interests. We consider RDF a good solution to publish the data extracted from accounts in machine-readable format as it facilitates explicit references to other (possible) representations of the accounts. If spreadsheets are employed, their use should be restricted to data input and calculations.

Presentations

This multi-speaker session reports on results from the first stage of a joint project of historians at the University of Regensburg, the Centre for Information Modeling - Austrian Centre for Digital Humanities at the University of Graz, and Wheaton College in Massachusetts. Short presentations from participants will highlight historical and technical features central to MEDEA.

Kathryn Tomasek will act as chair for the session and offer some brief observations about the advantages of international and cross-institutional collaborations for developing standards for semantically enhanced digital editions of accounts. In North America, the greatest support for producing digital editions of accounts tends to come from a small number of well-established documentary projects and from librarians and archivists who seek to encourage use of their sources. Resistance continues in the form of familiar objections emphasizing the labor-intensive nature of TEI and transcription generally. Both using transcription and markup in undergraduate classrooms and exploring opportunities for carefully curated crowdsourcing offer possible solutions. Graduate advisors in Asia and Europe seem to be more open to encouraging their doctoral students to explore digital edition of accounts than do those in the United States. Developing an international community of practice increases the likelihood that existing projects will seek ways to optimize their own data for the Semantic Web.

Kathrin Pindl and **Anna Paulina Orlowska** offer examples of ongoing interest in accounts as primary sources for historical study at the graduate level. Influenced by the work of British economic historian Richard C. Allen, Pindl hopes to develop better serial data through digital edition of accounts (Allen, 2001). Pindl will discuss her MEDEA working groups' experiences with beginning a digital scholarly edition of granary accounts from the St. Katharinenspitalarchiv in Regensburg. Orlowska will describe challenges to producing a digital edition of the accounts of Hanseatic merchant Johan Pyre, who was active in Danzig from 1421 to 1455. Since Pyre's bookkeeping methods developed through eight clear stages during those twenty-four years and contained unreliable temporal data, they demand a careful search for proper algorithms for their translation into digital media.

Oyvind Eide, **Cliff Anderson**, and **Georg Vogeler** will speak to some of the technical questions involved with data modeling for the Semantic Web. Eide will describe advantages of event-based modeling like that used in the CIDOC-CRM. Anderson will present observations based on his comparison of XML/TEI and the contemporary business tool XBRL-GL using information extracted from the Dutch periodicals *De Heraut* (*The Herald*) and *De Standaard* (*The Standard*), which were edited by the scholar, pastor, and prime minister Abraham Kuyper in the late nineteenth century.

Vogeler will close the presentations with a brief discussion of the value of Semantic Web technologies for historians interested in the "content" layer of accounts. RDFs/OWL allows modeling and encoding of the basic economic facts recorded in historical accounts and economic records. SKOS allows describing taxonomies of commodities, services, and monetary values recorded. They can be aligned into a common vocabulary. SPARQL allows aggregate querying of resources on these common facts together with individual data recorded. Digital editions of registers and accounts that not only publish text but try to express their interpretation of the text in a "content" layer and that publish this interpretation online with the help of semantic web technologies do what scholarly edition is meant to do: publish the critical analysis of the document by a competent scholar.

Funding

The MEDEA project is supported by an award from the Bilateral Digital Humanities program of the National Endowment for the Humanities and the German Research Foundation [grant numbers HG-229349-15]. Any views, findings, conclusions, or recommendations expressed in this session do not necessarily reflect those of the National Endowment for the Humanities or the Deutsche Forschungsgemeinschaft.

Bibliography

Allen, R. (2001). The Great Divergence in European Wages and Prices from the Middle Ages to the First World War, *Explorations in Economic History*, **38:** 411–47. doi: 10.1006/exeh.2001.0775.

Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M. (Eds.). (2011). Definition of the CIDOC Conceptual Reference Model. ISO: 21127:2014.

RDF Working Group. Resource Description Framework (RDF). http://www.w3.org/RDF/.

Sanderson, R. and Albritton, B. Shared Canvas Data Model. http://iiif.io/model/sharedcanvas/1.0/index.html.

Sarnowsky, J. (lead researcher), Die mittelalterlichen Schuldund Rechnungsbücher des Deutschen Ordens um 1400. http://www.schuredo.unihamburg.de/content/.

TEI Consortium. Guidelines for Electronic Text Encoding and Interchange. [Last updated on 15th October 2015]. http://www.teic.org/P5/.

Thaller, M. (2012). What is a text within the Digital Humanities, or some of them, at least? *Digital Humanities 2012: Conference Abstracts*. Hamburg: Hamburg University Press, July 2012, pp. 143-45.

Tomasek, K. and Bauman, S. (2013). Encoding Financial Records for Historical Research, *Journal of the Text Encoding Initiative*, **6**. doi: 10.4000/jtei.895.

Vogeler, G. (2014). Modelling digital edition of medieval and early modern accounting documents, *Digital Humanities 2014: Conference Abstracts*. Lausanne: EPFL and UNIL, July 2012, pp. 398-400.

W3C. OWL Web Ontology Language. http://www.w3.org/standards/techs/owl#w3c_all.

W3C. OpenAnnotation Data Model. http://www.openannotation.org/spec/core.

XBRL Global Ledger. https://www.xbrl.org/thestandard/what/globalledger/.

A Model for International Cooperation: Emblematica Online and Linked Data in Research and Pedagogy

Mara R. Wade
mwade@illinois.edu
University of Illinois, United States of America

Myung-Ja K. Han
mhan3@illinois.edu
University of Illinois, United States of America

Thomas Stäcker
staecker@hab.de
Herzog August Bibliothek, Wolfenbüttel

Pedagogy and Emblematica Online: Humanities Students in DH Projects

Mara R. Wade

Why DH a Pedagogy?

As Burdick, Drucker, Lunenfeld, Posner, and Schnapp state: "A new kind of digital humanist is emerging who combines in-depth training in a single humanistic subfield with a mix of skills drawn from design, computer science, media work, curatorial training, and library science" (Burdick et al, 166). Researchers with Emblematica Online took this challenge seriously, and determined ways to incorporate student researchers into our federally funded project as a means of investing in our shared DH future. Tier-one research institutions present particularly rich opportunities to advance the rapidly changing field of undergraduate research. Faculty members in the humanities at these universities have excellent opportunities to advance undergraduate research, and strong obligations do so. With graduate students, we can scaffold our mentoring to offer creative and productive experiences for DH pedagogy across all areas of the academic landscape. DH pedagogy involves experiential learning within a framework that emphasizes peer learning as well as more formal modes of instruction. Because DH pedagogy involves learning by doing, it emphasizes approaches in university teaching including both a vertical to a horizontal pedagogy.

The Emblem Scholars

This paper presents the research initiative with bachelor students at the University of Illinois, analyzes their research results, outlines the challenges, and makes suggestions for future efforts in DH pedagogy. "Digital Humanities: Emblematica Online" was a course for undergraduate researchers with no experience whatsoever in DH. Our goal was the long-term intellectual development of students with respect to the disciplines, concepts, workflows, and methodologies. This project also incorporated international research exchange and modeled experiential learning beyond the narrow confines of an undergraduate major.

The undergraduate research opportunity was a hybrid course, introducing students to early modern books and book history directly in the rich collections of the University Library, to literary criticism of texts and images with a tight focus on the European emblem, and to digital humanities concepts such as consistent vocabularies, multilingual thesaurus, best practices and standards for transcriptions of texts, and the importance of linked open data and semantic web technologies. The class integrated "a holistic learning by project approach" (Rehbein and Fritze).

The "Emblem Scholars" with six undergraduates was launched in spring semester 2013. The project PI organized the course as "Digital Humanities: Emblematica Online" and met twice per week. Project researchers introduced the virtual collection Emblematica Online and the workings of the Portal; they developed spreadsheets for motto transcriptions and the "stitching" program to collocate individual emblems with their metadata and create a URI from a handle server. The PI, together with a senior project consultant and a graduate assistant, directed these students. Students were also paired with faculty mentors from early modern studies who checked the students' transcriptions of emblem mottos from various early modern versions of European vernacular languages and Latin. These mentors also met with them once or twice during the semester about their research papers. The Emblem Scholars transcribed emblem mottos, associated this data with the emblems within their books, and wrote scholarly papers. Together they also presented two posters at the Undergraduate Research Symposium 2013 (see ill. 1). This pedagogy was largely student driven, and three students continued their research with Emblematica Online for several additional semesters. The continuing students created the website Emblem Scholars: Emblematica Online as an extension of their course work (see ill 2)

Student Outcomes and Next Steps

The student research associated with Emblematica Online is vital to training the next generation of DH scholars and users. All three continuing undergraduate students now plan graduate study in the humanities with DH integrated into their studies. Their research for Emblematica Online positions them ideally as users, and perhaps even creators, of future digital resources. There are further multipliers for positive outcomes. One student presented new aspects of her research at the Undergraduate Research Symposium 2014, while the two others studied abroad, in Ireland and France, respectively. During her semester abroad, the one student volunteered in the library, digitizing a photo album. She considers her research as an Emblem Scholar key to her acceptance to the highly competitive 2014 Andrew W. Mellon Summer Academy and Undergraduate Curatorial Fellowship Program at the Art Institute Chicago. One has an internship at a museum in the Chicago area, another has entered Library School.

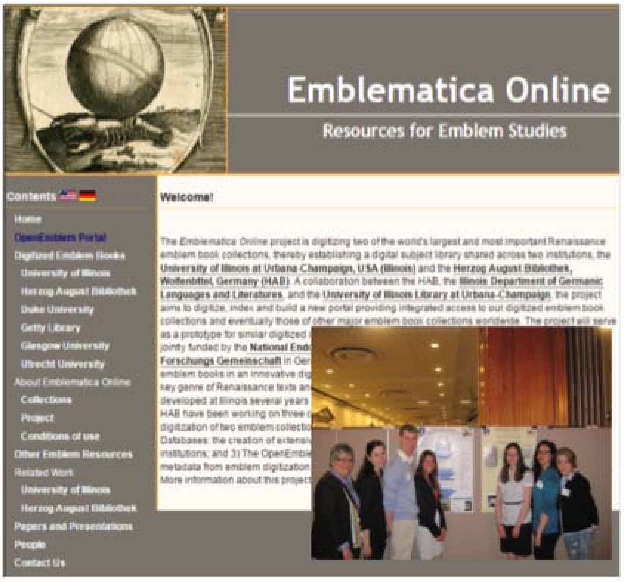
Challenges in Undergraduate Research

The first semester was time intensive to bring the students up to speed in all areas: DH, early modern literature and culture, working with rare books, and training in accurate transcription of foreign languages. (The six students came from Spanish, Communications, Mathematics, Psychology, English, and Art History.) Taking ownership of their work, responsibility for accuracy in all workflows, and dependability in meeting deadlines were habits they quickly developed. These students were smart and eager to work; they developed the necessary skills rapidly. We received a competitively awarded start-up fund of $1,500 for a new initiative in undergraduate research. That contributed to the purchase of materials (a book, a large-capacity flash drive, some computer hardware) and to printing posters. All teaching was done as an overload to the normal courses assignments. Regularizing funding and integrating the teaching of such courses into the normal operations of an academic department are clear desiderata. Owing to generous consortium and renewed SLCL funding, the three students will participate in a DFG funded event at the Newberry Library, Chicago, "Emblematica Online Workshop: Link Open Data – Developing an Ontology for Annotating Emblems."

DH and Critical Digital Pedagogy

We posit that investment in undergraduate (bachelor) education is well worth the time and resources spent. The "Emblem Scholars" participated in aggregated, distributed, collaborative, and open learning techniques (Gold, 2012a) within the framework of a federally funded project, Emblematica Online, that offers an ideal test bed for new knowledge, best practices, and emerging standards for all researchers involved in the project. Significantly, it has introduced students at all level of study in the humanities and library science to humanistic digital research. The researchers at Emblematica Online invested in their long-term trajectory of scholarly and intellectual development. We also developed a kind of "critical digital pedagogy" that educates for the long run through intensive mentoring and experiential learning. As Sean Michael Morris and Jesse Strommel define it: "Critical Pedagogy is an approach to teaching and learning predicated on fostering agency and empowering learners… ." The experience of Emblematica Online suggests that DH pedagogy is praxis-oriented teaching that positions students to learn experientially in a hybrid setting. While the usual project "tasks" are certainly part of the experience, the focus is on the development of a theoretical and conceptual understanding of the intersection of traditional and digital humanities. By situating learning directly in the research experience, students acquired a broad range of skills and concepts that exceeds a narrow academic focus. These students see DH as an integral part of the way we do humanities (McCarty, 24). Emblematica Online is a scholar driven project, and this synergy is reflected in the pedagogical outcomes.

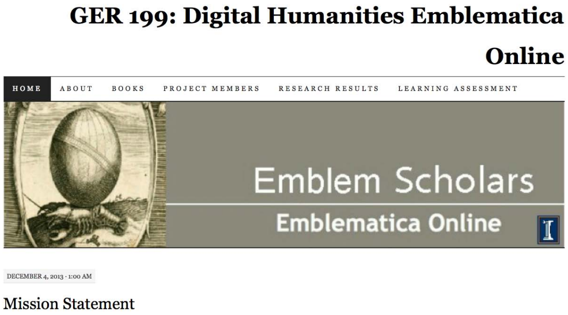
In her blog post, "Commit to DH people, not DH Projects," Miriam Posner, University of California, Los Angeles, makes the case for investing in the human resources clustered around projects, such as Emblematica Online. Posner raises the questions: "What if you saw that training period as an investment in healthy, long-lasting relationships? What if we saw digital humanities as a long-term investment in scholarly growth, not a short-term investment in projects?" The student researchers with Emblematica Online learned a range of critical thinking, technical, intellectual, and administrative skills that are the most transferrable and will serve them well in the future. They are well prepared to be DH citizens.



Welcome!

The *Emblematica Online* project is digitizing two of the world's largest and most important Renaissance emblem book collections, thereby establishing a digital subject library shared across two institutions, the **University of Illinois at Urbana-Champaign, USA (Illinois)** and the **Herzog August Bibliothek, Wolfenbüttel, Germany (HAB)**. A collaboration between the HAB, the **Illinois Department of Germanic Languages and Literatures**, and the **University of Illinois Library at Urbana-Champaign**, the project aims to digitize, index and build a new portal providing integrated access to our digitized emblem book collections and eventually those of other major emblem book collections worldwide. The project will serve as a prototype for similar digitized
jointly funded by the **National End**
Forschungs Gemeinschaft in Ger
emblem books in an innovative dig
key genre of Renaissance texts an
developed at Illinois several years
HAB have been working on three
digitization of two emblem collectio
Databases: the creation of extensi
institutions; and 3) The OpenEmbl
metadata from emblem digitization
More information about this project

Illustration 1: Undergraduate Researchers Presenting at the University of Illinois 2013



GER 199: Digital Humanities Emblematica Online

DECEMBER 4, 2013 · 1:00 AM

Mission Statement

Illustration 2: Undergraduate Researchers' Website

Transforming SPINE metadata to Linked Open Data

Myung-Ja K. Han
Timothy W. Cole
Maria Janina Sarol
Patricia Lampron

Background

Emblem books flourished in Europe as a popular literary genre from 1531, the publication date of the first emblem book, through the mid-18th century. The form of the emblem book is compound, integrating text and graphics, and highly contextual, often influenced by cotemporaneous events. Remaining copies of printed emblem books are widely dispersed in libraries across Europe and North America today. To facilitate access and better support Emblem Studies, a large segment of the corpus has now been digitized and made discoverable through the Emblematica Online Portal (http://emblematica.library.illinois.edu/). Because the individual emblems within emblem books are themselves of scholarly interest, emblem books pose interesting descriptive challenges for libraries. A domain-specific, XML-based metadata schema was developed and implemented to address these challenges and support a measure of interoperability within the Emblem Studies community (http://diglib.hab.de/rules/schema/emblem/emblem-1-2.xsd). However, emblems are of interest to scholars in additional disciplines, e.g., art historians, historians of Renaissance and Baroque cultures, comparative literary scholars, etc. To maximize utility it is crucial to integrate digitized emblem resources with other kinds of resources used by scholars in multiple domains. As a step towards this vision, this presentation will report on work done to transform our emblem-specific XML metadata into Linked Open Data in order to support broader discovery and access and to better integrate digitized emblem resources with other resources on the Web.

The SPINE Metadata Schema and Discovery Services

Discovery and browse services provided through the Emblematica Online Portal are possible because of the emblem-specific SPINE metadata schema. The genesis of the SPINE schema was an international workshop on the digitization of emblems held in 2001 at the Centre for Emblem Studies at the University of Glasgow. The participants in the workshop, drawn from widely dispersed institutions, recognized the need for a shared metadata vocabulary to support interoperability across emblem collections. The classes and properties of the SPINE schema were subsequently documented by Stephen Rawles of the University of Glasgow (2004). Thomas Stäcker at the Herzog August Bibliothek Wolfenbüttel implemented the schema proposed by Rawles in XML. To express most book-level properties, Stäcker integrated the Metadata Object Description Schema (MODS) into SPINE. He supplemented MODS classes and properties with emblem-level descriptive classes and properties and added a few additional book-level properties to record the provenance of digitized emblem volumes, i.e., to link digitized items back to the print item from which they are derived. SPINE is now available as an XML based metadata schema, currently in version 1.2 (Stäcker, 2012).

As implemented in the Emblematica Online Portal, the SPINE metadata schema, designed for describing emblem books, the emblems contained in those books, and copy-specific information about the book, allows digital humanities scholars access to digitized emblem resources at any of three different levels granularity, an entire emblem book, an individual emblem, and an individual pictura, while providing detailed descriptive information at each level (Cole et al. 2012). Iconclass, a multilingual, extensible classification system for visual cultural content containing more than 28,000 hierarchically ordered descriptors, has been used extensively by scholars and curators to describe emblem picturae. Most recently, the team has exploited Iconclass.org linked open data services ("Iconclass as Linked Data" 2015) to facilitate emblem discovery. By using the available Iconclass Web services, the Emblematica Online Portal can now support multilingual search and browsing services for Iconclass headings as well as browsing broader and narrower Iconclass headings, demonstrating the benefits of being a linked open data consumer (Cole et al. 2013).

Exposing SPINE Metadata to the Web

As a next step, the Emblematica Online project team has been exploring ways to become a linked open data producer by publishing SPINE metadata as linked open data. To accomplish this, the Emblematica Online project team takes a rather unique approach. Instead of creating a new ontology for the SPINE metadata schema, the project team is focusing on discovery and visibility of digitized emblem resources on the web, in other words, using linked data as a vehicle that brings the emblem resource information contained in the SPINE metadata to the web. To make SPINE metadata more web friendly, i.e., indexed and searchable by web search engines including Google, and compliant with current linked open data work being conducted across cultural heritage institutions, the Emblematica Online project team has developed a workflow that transforms SPINE metadata to Resource Description Framework in Attributes (RDFa) based web pages for the presentation of both emblem books and emblems. As an extension to HTML5, RDFa allows web search engines to generate better search results through the attributes embedded in the web pages, ultimately improving the visibility of these resources on the web. The RDFa attributes are being populated relying on schema.org semantics, since these semantics are recognized and used by major web search engines as a de facto linked data markup standard, whenever schema.org semantic meanings aligned well with the SPINE metadata elements. When there are no corresponding semantics in schema.org, or it is necessary to keep the semantic meanings of the SPINE schema, schema.org extensions (2015) are employed to represent emblem specific information, such as Pictura, Iconclass, subscriptio, motiv, emblemParts, and so on. The figure below shows how the SPINE XML description of an emblem pictura that includes an Iconclass descriptor (figure 1-a) can be transformed to RDFa with semantics from schema.org (figure 1-b).

```
<pictura xml:id="E001352_P1" xlink:href="http://djatoka.grainger.illinois.edu/...">
<iconclass rdf:about="http://www.iconclass.org/rdk/25F711(GRASS-HOPPER)(+45)">
<skos:notation>25F711(GRASS-HOPPER)(+45)</skos:notation>
</iconclass>
```

```
</pictura>
```

Figure 1-a: < pictura> element in a SPINE XML metadata that can include one or more iconclass descriptors.

```
<div rel="sc:hasPart" typeOf="emb:Pictura">
<a property="sc:url" href="http://djatoka.grainger.illinois.edu/...">Pictura Image</a>
<div rel="sc:about" typeOf="sc:CpnceptCode">
<span rel="sc:codingSystem" resource="http://iconclass/org/"/>
<span property="sc:name">inscets: grass-hopper (+ animals eating and drinking)</span>
<a property="sc:sameAs" href="http://www.iconclass.org/rdk/25F711(GRASS-HOPPER)(+45)">
<span property="sc:codingValue">25F711(GRASS-HOPPER)(+45)</span>
</a>
</div>
</div>
```

Figure 1-b: The same <pictura> element transformed to RDFa using schema.org semantics and schema.org extension with Emblem specific types (classes) like emb:Pictura as well as mini-skos.

Conclusion

Early experimentations with linked data undertaken by the Emblematica Online Portal have allowed emblem scholars to access related resources available in the Virtual Printroom (Virtuelle Kupferstichkabinett, VKK), jointly managed by the Herzog August Bibliothek and the Herzog Anton Ulrich Museum in Germany, and Festkultur Online, developed and maintained at the Herzog August Bibliothek, since all three sites use Iconclass as a descriptive vocabulary for their resources. Additionally, exposing SPINE metadata in RDFa based HTML pages would expand relationships between information resources in SPINE and other contextual information available on the web that is represented with the same schema.org semantics, such as information about the author or historical or cultural events related to emblem books and emblems, ultimately enabling emblem scholars to access additional information in conjunction with emblem resources available at the Emblematica Online Portal. This presentation will discuss why Emblematica Online experimented with linked open data, describe details of the workflow, challenges, and lessons learned from this SPINE to linked open data transformation work, and enumerate future plans for using additional linked open data sources to improve user experiences.

Linked Open Data for Emblem Research

Dr. Thomas Stäcker

This paper presents a project carried out in close collaboration with the current research of Emblematica Online, University of Illinois (Prof. Mara Wade, PI). It aims to strengthen international co-operation in the field of emblem studies, to jointly develop and apply technical models, and to make use of synergies to establish transnational digital collections and data pools that can be freely used by the international emblem community. The new project at the Herzog August Bibliothek, Wolfenbüttel, proposes to develop and integrate concepts and models that can be applied to and employed by other areas within the burgeoning field of Digital Humanities. It is divided into four distinct but interconnected areas of activity. First, the data in the collection itself will be enlarged; secondly, emblem books will be transcribed and made available as full text; thirdly, and that is the most innovative part of the project, data and metadata will be modeled and designed in such a way that they can be used in the semantic web; and finally, interfaces and web services will be enhanced or established to facilitate the harvesting and retrieval of emblem data.

A central feature of this project is the development of an ontology that allow scholars to analyze emblems by offering a vocabulary that is apt to ascertain identities, resemblances, or significant differences of emblems digitized at various places on the web. Supported by suitable tools scholars will ideally be able to encode in RDF (see example, Ill. 1) that there is an identity with respect to meaning by relating a helmet with bees to a camel walking through the eye of a needle (both of them allude to the topic peace, see Ill. 2), or they will be able to assemble identical emblems from various books or will characterize instances where the same motto is used for various picturae. The project draws on a comprehensive corpus of emblems that were digitized in several larger projects within the last decade. It is meant to move a step forward in that it no longer intends to make source material merely available, but attempts to offer means and methods for practical research by applying semantic web techniques and standards such as the Open Annotation Collaboration and Iconclass' RDF representation.

```
# resemblance that cannot be further determined
emblem:resemblance rdf:type skos:Concept;
skos:prefLabel “Ähnlichkeit oder Gleichheit”@de;
skos:prefLabel “resemblance”@en;
skos:definition „Nicht weiter bestimmte Ähnlich- oder Gleichkeit“@de.
#subordinated concepts of similarity and resemblance
emblem:sameness rdf:type skos:Concept;
skos:prefLabel “Gleichheit”@de;
skos:prefLabel “sameness”@en;
skos:narrower emblem:resemblance.
emblem:similarity rdf:type skos:Concept;
skos:prefLabel “Ähnlichkeit”@de;
skos:prefLabel “similarity”@en;
skos:narrower emblem:resemblance.
```

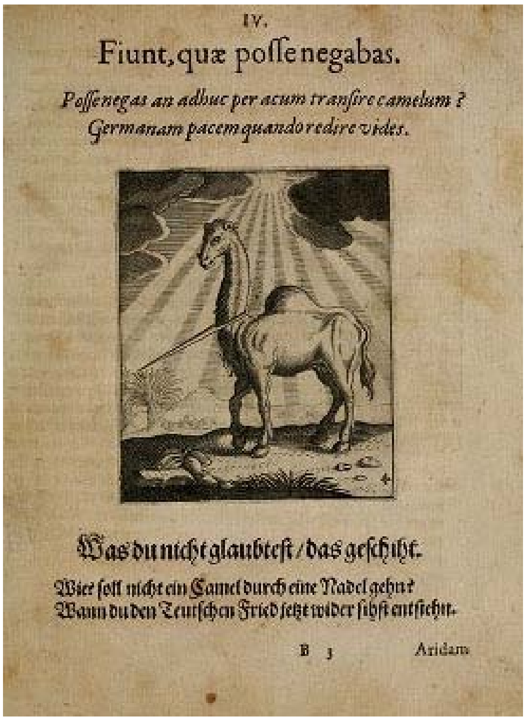
#causality in a most general sense, e.g. also „exemplar for", „cause for " etc.

emblem:causality rdf:type skos:Concept;
skos:prefLabel "Verursachung"@de;
skos:prefLabel "causality"@en.

#contiguity, e.g. emblems in a one book, emblems occuring in the same period of time etc.

emblem:contiguity rdf:type skos:Concept;
skos:prefLabel "Räumliche oder zeitliche Nähe"@de;
skos:prefLabel "contiguity" @en
Usw

Ill. 1: URI-Emblem-1d http://hdl.handle.net/10111/EmblemRegistry:E000004



IV.

Fiunt, quæ poſſe negabas.

Poſſe negas an adhuc per acum tranſire camelum?
Germanam pacem quando redire vides.

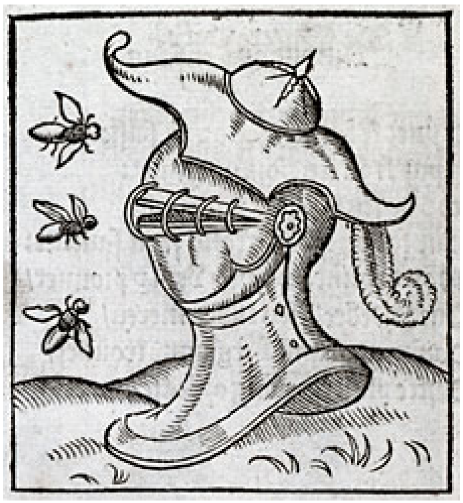
Was du nicht glaubteſt / das geſchiht.

Wie ſoll nicht ein Camel durch eine Nadel gehn?
Wann du den Teutſchen Fried ſetzt wider ſihſt entſtehn.

B 3 Aridam

<URI-Emblem-1> emblem:sameTopic <URI-Emblem-2>

URI-Emblem-2 http://www.emblems.arts.gla.ac.uk/french/emblem.php?id=FALa045.



III.2

External links

- Arkyves: http://arkyves.org/
- Emblematica Online: http://emblematica.library.illinois.edu/
- Emblem Scholars: http://emblematicaonlineuiuc.wordpress.com/
- Festkultur Online: http://www.hab.de/en/home/research/projects/festkultur-online.html
- Hybrid Pedagogy: http://www.hybridpedagogy.com/journal/critical-digital-pedagogy-definition/
- Mellon Summer Academy, Art Institute Chicago :http://www.artic.edu/employment/andrew-w-mellon-summer-academy-and-undergraduate-curatorial-fellowship-program
- Mini-skos proposal for schema.org: https://dvcs.w3.org/hg/webschema/file/default/schema.org/ext/miniskos.html
- Open Annotation Collaboration: http://www.openannotation.org/
- Posner, Miriam, Blog: http://miriamposner.com/blog/commit-to-dh-people-not-dh-projects/#more1687
- RDFa: https://rdfa.info/
- Schama.org: http://schema.org/
- School for Literatures, Cultures, and Linguistics, University of Illinois: http://www.slcl.illinois.edu/
- Virtual Printroom (Virtuelle Kupferstichkabinett, VKK): http://www.virtuelles-kupferstichkabinett.de/

Bibliography

Alexander, Bryan and Rebecca Frost Davis. (2012). Should Liberal Arts Campuses Do Digital Humanities? Process and Products in the Small College World, Debates in the Digital Humanities, ed. Matthew Gold. Minnesota: U Minnesota Press, pp. 368-389.

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Posner, and Jeffrey Schnapp. (2012) Digital Humanities. Cambridge, MA: MIT Press.

Cole, Timothy W. and Myung-Ja Han. (2011). The Open Annotation Collaboration Phase I: Towards a Shared, Interoperable Data Model for Scholarly Annotation, Journal of the Chicago Colloquium on Digital Humanities and Computer Science, 1. 3.

Cole, Timothy W. and Myung-Ja Han. (2012). Developing Digital Library Services to Support Emblem Studies, Emblematica. A Journal for Emblem Studies, 19, pp. 65-94.

Cole, Timothy W., Myung-Ja Han, Mara Wade, and Thomas Staecker. (2013). "Linked Open Data & the OpenEmblem Portal." Digital Humanities 2013. Available at http://dh2013.unl.edu/abstracts/ab-359.html.

Cole, Timothy W., Myung-Ja Han, Jordan Vannoy. (2012). "Descriptive Metadata, Iconclass, and Digitized Emblem Literature." Proceedings of the 12th Annual Joint Conference on Digital Libraries. New York: Association for Computing Machinery. pp.111-120.

"Extension Mechanism." 2015. Acceded on October 30 2015. Available at https://schema.org/docs/extension.html.

Gold, Matthew. (2012a). Looking for Whitman: A Multi-Campus Experiment in Digital Pedagogy, Digital Humanities Pedagogy: Practices, Principles and Politics, ed. Brett D. Hirsch. Openbook, pp. 151-176.

Gold, Matthew. (2012b). Debates in the Digital Humanities. Minnesota: U Minnesota Press.

Hirsch, Brett D. ed. (2012). Digital Humanities Pedagogy: Practices, Principles and Politics, Openbook. http://www.openbookpublishers.com/reader/161#page/1/mode/2up

"Iconclass as Linked Data." Last edited on September 18 2015, accessed on October 30 2015. Available at http://www.iconclass.org/help/lod.

McCarty, Willard. (2012). The Residue of Uniqueness, Controversies Around the Digital Humanities, ed. Manfred Thaller. Historical Social Research, 37.2, pp. 24-45, here 24.

Rawles, Stephen. 2004. "A SPINE of Information Headings for Emblem-Related Electronic Resources." In Digital Collections and the Management of Knowledge: Renaissance Emblem Literature as a Case Study for the Digitization of Rare Texts and Images, edited by Mara Wade, 19-28. Salzburg: DigiCULT.

Rehbein, Malte and Christiane Fritze/ (2012). Hands-On Teaching Digital Humanities, Digital Humanities Pedagogy: Practices, Principles and Politics, ed. Brett D. Hirsch. Openbook, pp. 47-78.

Reid, Alexander. (2012). Graduate Education and the Ethics of Digital Humanities, Debates in the Digital Humanities, ed. Matthew Gold. Minnesota: U Minnesota Press, pp. 350-367.

Stäcker, Thomas. 2012. "Practical Issues of the Wolfenbüttel Emblem Schema." In Emblem Digitization: Conducting Digital Research with Renaissance Texts and Images, edited by Mara R. Wade. Early Modern Literary Studies Special Issue 20. Accessed 27 May 2015, http://purl.oclc.org/emls/si-20/WADE_Staecker_EMLS_Schema.htm>

Thaller, Manfred, ed. (2012). Controversies Around the Digital Humanities, Historical Social Research, 37.2

When DH Meets Law: Problems, Solutions, Perspectives

Andreas Witt
witt@ids-mannheim.de
IDS Mannheim, Germany

Paweł Kamocki
pawel.kamocki@gmail.com
IDS Mannheim, Germany; Université Paris Descartes, France;
Westfälische Wilhelms-Universität Münster, Germany

Digital data is fuel for data-intensive science. Access, re-use and sharing of this data, however, while required by academic ethos and good practice, are often highly restricted by legal frameworks. In particular, the areas of law that can affect research data are: intellectual property (copyright and database rights) and personal data protection. These issues are particularly relevant in the field of Digital Humanities, which study various aspects of human activities in general, and their creative and social aspects in particular. In fact, most of the research data in digital humanities are within the scope of either intellectual property or data protection law, which means that they cannot be freely accessed, re-used and shared without a permission of the right holder or the data subject's consent.

Moreover, research funding agencies require more and more often that the results (and underlying data) of research projects that they fund be made available in Open Access. Open sharing of research data and outcomes is often perceived nowadays as an ethical obligation in contemporary science, but it cannot be done in a satisfactory way without addressing legal concerns (such as appropriate licensing and rights clearance).

Legal issues are increasingly being taken into account in the preparation phases of many research projects. Scientists who do not consider legal issues in their research activities may be exposed to certain legal risks. Existing statutory exceptions for research rarely provide for enough relief (even though lobbying efforts are being made to extend their scope). In short, modern science in general, and Digital Humanities in particular, are more concerned with legal issues than ever before.

The purpose of the multiple paper session we propose is to emphasize the problem, discuss various technological and organizational solutions, as well as future legal challenges that DH researches will have to face.

Three papers by authors with both legal training and hands-on experience with D-SSH research data management will be presented. The first one compares organizational and technical solutions adopted in the field of DH and Social Sciences. The second discusses research data licensing, and presents existing tools that help researchers in the process. The third paper examines legal and ethical aspects of stylometry and authorship attribution research.

"One Does Not Simply Share Data". Organisational and Technical Remedies to Legal Constraints in Research Data Sharing – building bridges between Digital Humanities and the Social Sciences

Pawel Kamocki
Katharina Kinder-Kurlanda
Marc Kupietz

Within the Social Sciences there exists a long tradition of data sharing, which is facilitated by infrastructure institutions such as the GESIS Data Archive that has been providing survey data to researchers since the 1960s. More recently various technical solutions aiming to grant secure and user-friendly access to data requiring special protection have been emerging in the field.

The DH also have a well-established tradition of research data sharing. In linguistics, for example, digital text collections have also been published since the 60s (e.g. the Brown Corpus or the Mannheimer Korpus). First software solutions to share the data and to make it accessible to other researchers emerged in the late 1980s and started to boom with the appearance of the WWW in the early 1990s.

Legal barriers to data sharing

Legal issues have long been identified as barriers to research data sharing. They can be divided into two categories: those related to intellectual property rights and those related to privacy laws.

Intellectual property rights — such as copyright and the database right — grant the rights holders certain exclusive rights (monopolies), i.e., rights to exclude others from the use of their property. For example, in order to copy and distribute a copyright-protected work or a database, one normally has to obtain permission from the rights holder, usually in an agreement known as a license. This highly affects DH — disciplines fueled by digital data issued from human creative activities, which normally qualify for copyright protection.

It is essential to understand that intellectual property is similar to "traditional" (i.e. corporeal) property. Therefore, it can be said that most research data in DH in fact belong to a third party (author or publisher). The right to property is a fundamental freedom, which overrides freedom of research. As a consequence, statutory research exceptions are rarely enough to allow use of copyright-protected data in research projects.

Researchers in DH therefore need to obtain licenses for the use of data, which is not an easy task. The negotiations may be time-consuming, and the result is not always satisfactory. In practice, licenses signed with e.g. publishers are often very restrictive and non-transferable.

Another legal framework that affects researchers in areas such as medicine or the social sciences, but also in the DH, is personal data protection. In principle, personal data (i.e. any information related to an identifiable person) can only be processed if the data subject has validly consented to the processing. While it is true that anonymised data can be freely processed, anonymisation may strip a dataset from most (if not all) of its informational and scientific value. The obligation to obtain consent is particularly burdensome when it comes to older data that has been collected without consent, or that has been collected for a different purpose (re-purposing normally necessitates a new consent). Also, in practice, consent rarely covers transfer and sharing of data with other researchers.

Social Science approach

In the Social Sciences quantitative survey data is particularly interesting for sharing. The highly controlled and well-documented ways of gathering data in large-scale survey programmes make the data highly reusable in a methodologically sound way. The GESIS Data Archive for the Social Sciences in Germany provides survey data for secondary use and thus allows researchers to share collected data in a user-friendly, searchable and standardised manner. Due to data protection legislation participants of survey data provided by the archive must not be re-identifiable, or only with a disproportionate amount of time, expense and labour. Anonymization challenges usually occur once detailed geographical as well as demographical information has been collected.

To improve data sharing several solutions have been found for the Social Sciences. For example, most data at the GESIS archive is anonymised and thus can be provided for download via the online data catalogue. Some datasets containing more detailed information are provided employing secure data access solutions. A combination of contractual, organizational and technical safeguards is employed to ensure that individuals' rights to anonymity are protected. For example, for particularly disclosive data, researchers must visit a safe room where a completely encapsulated virtual research environment is provided via a thin client. They cannot download any data or access the internet. They are also not allowed to bring mobile phones or other electronic devices. Any analysis output they produce is intellectually assessed for its level of disclosiveness and only handed to researchers once the output criteria are fulfilled. Only users whose signed usage agreements (detailing the research topic and the methods applied) have been approved can use the safe room.

Secure remote solutions either provide researchers with a secure connection to an encapsulated work environment as described above or allow the submission of code and syntax to be run on the data by the data provider. All remote solutions need to be secured from threats posed by using the internet.

DH approach

To cope with legal challenges, to make research data as openly accessible as possible, and to enable traceability and replicability without interfering with legitimate interests of rights holders, the disciplines that deal with language as their primary research data, particularly linguistics, have developed several strategies.

As already discussed above, usually the only way to acquire text for research purposes is to obtain licenses from copyright holders. The copyright holders can sometimes be convinced to provide scientific licenses for free or for comparatively low fees as long as they do not interfere with the company's business model. Given that some institute is willing to conclude partially transferable license agreements with rights holders and license agreements with end-users, thus acting as an intermediary between both parties, and to provide software that enforces the license restrictions and provides all retrieval and analysis functions that researchers need, every group's interests can be satisfied. Indeed, this model has worked successfully at the Institute of German Language (IDS) since the beginning of the 90s and also for most other providers of national and reference corpora.

The problem with the intermediary model alone is that it requires the intermediary to provide all functions that are required for any researcher. While, for example, the requirements for more traditional linguists could be satisfied, this was not possible for researchers in the field of text mining and computational linguistics where the methods of analysis themselves are in the center of research and therefore subject to rapid change.

To meet the needs of such user groups, another idea, recently described for linguistics but traditionally used in data-intensive disciplines such as climate research, has to be put into action. Extending Gray's (2003) famous claim "put the computation near the data" to situations where the data cannot be moved due to license restrictions, data providers also provide mechanisms that allow end-users to run their analysis software on the data located at the provider, making sure that the software does not violate any license restrictions.

Another remedy that is often applied in the context of corpora that are based solely on texts from the WWW, is not to share the research data itself, as this could be a copyright violation, but rather to share the software that retrieves the texts from the Web. Depending on the the application scenario and the local legislation, this technique can enable end-users to benefit from statutory exceptions. An unwanted side-effect of this approach is that the identity of the corpus data retrieved by different runs of the retrieval software cannot be guaranteed. However, there is some consensus in the research community that the sights with respect to demands on replicability and persistency of research data necessarily have to be lowered to a realistic standard that takes into account legal restrictions. This aspect has also recently entered the best-practice guidelines of the German Research Foundation.

Conclusions

D-SSH are dealing with data surrounded by legal issues. Traditionally, Social Science researchers process more privacy-sensitive information, whereas DH researchers work with data protected by intellectual property rights, usually belonging to third parties; the division, however, is not clear-cut. Some disciplines in DH (such as linguistics) also have to deal with privacy-sensitive material, and the Social Sciences are concerned if not by copyright, then by other branches of intellectual property (such as database right). Increasingly commercial data owners such as social media companies are becoming important.

Institutes in both disciplines have developed idiosyncratic ways of coping with legal restrictions which provide satisfactory results. This shows us that some legal issues may be resolved by appropriate organisational, technical and infrastructural solutions; the comparison between DH and Social Sciences, however, demonstrates that both disciplines still have room for improvement and that there is a lot that they can learn from each other.

«Trust me. I'm a License Selector». Licensing for Digital Humanities

Paweł Kamocki & Pavel Stranak

Lack of legal interoperability (i.e. a situation in which a dataset cannot be used due to incompatible licensing restrictions on its various parts) has been identified as one of the major obstacles for data access, sharing and re-use. This is particularly relevant in the field of Digital Humanities, where data are often protected by copyright (i.e., they are created by human authors). While it is true that some data (e.g. those obtained from press editors) are only available to researchers under very restrictive license agreements, in fact quite often legal interoperability problems can be solved by proper licensing of research outcomes. Indeed, despite the fact that openness and reproducibility of results have long been identified as cornerstones of the scientific community, in practice many digital datasets and tools are being shared under licenses that are unnecessarily restrictive or not fit for the purpose, or even without any licenses at all. This is probably due to the fact that the task of choosing an appropriate license may seem difficult for an average researcher with a limited access to legal advice. As a response to that problem, attempts have been made to build tools (referred to as License Choosers, License Selectors or even License Wizards) that would guide the users through the jungle of available public licenses and allow him to choose one that is the most suitable for his needs.

Before these License Selectors can be presented and assessed, it is essential to define the notion of a public license. A public license is a license that grants certain rights not to an individual user, but to the general public (every potential user). Public licenses for software has been known since 1980s (when software licenses such as BSDL, MIT or GNU GPL emerged). However, public licenses for other categories of works (including datasets) only appeared in the 21st century, mostly due to the creation of the Creative Commons foundation. The latest version of the CC license suit (including six licenses, a waiver and a public domain mark), CC 4.0, is well adapted for datasets, as it covers not only copyright, but also the sui generis database right, but older versions are still in use. While choosing a license, one has to keep in mind that the licenses which are appropriate for software are not appropriate for data and vice versa. Moreover, not all public licenses are 'open', i.e. not all of them meet the requirements for Open Access/Open Data/Open Source label. In our paper, we would like to briefly demonstrate three online tools made specifically for licensing of research material.

The Licentia tool (http://licentia.inria.fr/visualize) has been developed in 2014 by Cardellino for INRIA (French Institute for Research in Computer Science and Automation) is in fact a conglomerate of three tools: a License Search Engine (which allows to identify licenses that meet a set of requirements defined by the user), a License Compatibility Checker (which assesses whether two licenses are compatible, i.e. whether material licensed under those two licenses can be 'mixed') and a License Visualiser (an interesting extra feature which produces graph-based visualisations of licenses expressed in ODRL - Open Digital Rights Language Deontology).

The ELRA (European Language Resources Association) License Wizard (http:// wizard.elda.org), released in April 2015, allows users to define a set of features and browse corresponding licenses. For now, the tool only includes CC, META-SHARE and ELRA licenses, so it is particularly useful for language resources.

Finaly, the Public License Selector (http://ufal.github.io/public-license-selector/) developed by Kamocki, Stranak and Sedlak in 2014 as a cooperation between two CLARIN centres (IDS Mannheim and Charles University in Prague) uses an algorithm (a series of yes/ no questions) to assist the user in the licensing process. It allows to choose licenses for both data and software, and features a built-in License Interoperability Tool. Licenses that meet the 'open' requirement are clearly marked. Finally, unlike the two other tools, it is made available under Open Software/Open Data conditions.

All of these tools have both advantages and disadvantages; their biggest disadvantage is that they use (to a different degree) a very specific language, which in fact requires basic knowledge of Intellectual Property Law from the user. They also necessarily involve a certain degree of over- or undergeneralization, especially when it comes to assessing license interoperability. Nevertheless, they remain very useful for the research community and may indeed help facilitate re-use and sharing of tools and data in Digital Humanities.

Legal and Ethical Aspects of Authorship Attribution Using Stylometry - EU and US Perspectives

Erik Ketzan & Paweł Kamocki

Introduction

Authors have written anonymously since the invention of writing, and the growing digital humanities field known variously as stylometry / computational stylistics / authorship attribution often aims to discover the identify (or rule out the identity) of anonymous authors.

Depending on whether such authors are living, whether the works in question are protected by copyright, and what the aims of the digital humanities research is, vastly different legal frameworks govern such research in the European Union and United States. The strong data protection laws of the EU seem to prohibit certain types of authorship attribution research, while researchers in the US have vastly fewer restrictions regarding data protection regulations. In addition to data protection, the acts of copying and analyzing texts for the purposes of stylometry raise copyright concerns. In the US, these acts seem to be largely allowed by the fair use doctrine. In the EU, these fall into more questionable legal territory, although new laws regarding text and data mining offer improved guidance to researchers.

As laws concerning research in the digital age are being revisited in both the US and EU, it is important to see where stylometry falls under current legal frameworks, and how, and whether, researchers should advocate for changes to law. Finally, we argue that a parallel debate regarding the ethics of stylometric research should be begun. As stylometric research and technology continues to improve, with promises of improved reliability of authorship attribution, researchers should begin to debate which questions researchers should ethically tackle, not only which questions they can.

Stylometry of anonymous authors under EU law

In the EU, where the memory of totalitarian governments is still present, Member States value privacy very highly. This is translated in the legislation, where the right to be and remain anonymous is not only protected by rules on the processing of personal data, but sometimes also to an extent guaranteed by copyright laws.

The Data Protection Directive is the primary source of laws governing the processing of personal data, and guides Member States in protecting "the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data." The Directive defines personal data as, "any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity." Processing of personal data is defined as, "any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use," etc.

In general, processing of personal data can only be done if the data subject (i.e. the person that the data refer to) has unambiguously given his consent. Moreover, the Data Protection Working Party (a body composed of representatives of National Data Protection Authorities from each Member State and whose purpose is to give expert advice on the interpretation of the Data Protection Directive) clearly stated that information that does not relate to an identified person, but is collected for the purpose of identification, shall also be regarded as personal data.

European researchers engaging in stylometric research for the purpose of identifying a living author therefore engage in the processing of personal data, and are subject to the rules and restrictions of the Data Protection Directive and related Member State data protection laws. Alternatively, consent could be obtained to process the personal data, but this leads to the absurd suggestion that researchers obtain permission from an anonymous author so that they can guess at his/her identity.

While the current framework allows for alternative grounds for lawfulness of processing (other than consent), such as e.g. pursuit of legitimate interests, these provisions remain vague and do not guarantee the necessary legal security for researchers. Exceptions from the rules set up by the Directive exist, but only cover very special cases such as freedom of journalistic and artistic expression, public security or (to a limited extent) historical, statistical and scientific research.

An inevitable conclusion, however, is that Personal Data Protection law in the EU protects anonymous authors from being identified against their will, at least when they are still alive.

Anonymity of authors is also addressed by many national laws on copyright. Although anonymous works benefit from a significantly shorter term of protection (70 years after publication, and not 70 years after the death of the author), the anonymity of the author is nevertheless protected; his rights can be exercised by a proxy (usually an agent or publisher). Moreover, in some jurisdictions (i.e., in France), inaccurate attribution of authorship can be regarded as violation of moral rights (i.e., a form of copyright infringement) of both the real author and the falsely attributed one.

Stylometry of anonymous authors under US law

The legal framework of the United States governing stylometry of anonymous authors is vastly different from the EU. The United States has no single general data protection law. The First Amendment of the United States Constitution guarantees the right to free speech, and a broad right to privacy has been inferred from the Constitution by the United States Supreme Court. A number of state constitutions, such as California, explicitly mention privacy as well.

Courts in the US have recognized certain rights to anonymity, most notably in McIntyre v. Ohio Elections Commission, 514 U.S. 334 (1995), where the Supreme Court held that the freedom to publish anonymously is protected by the First Amendment, and extends beyond the literary realm to the advocacy of political causes. Whether such a right extends to a researcher attempting to remove that anonymity is an open question.

Regarding copyright, there are strong arguments that the acts of copying and data mining text for research purposes are covered by the fair use doctrine, especially after the landmark Google Books case, which held that the scanning of books and making snippets available in search engines is a fair use. As a typical stylometric analysis involves the copying of texts and analysis on a single computer, without distribution of snippets (in other words, infringing less upon exclusive rights than the facts in Google Books), the acts seem to be covered by the fair use framework.

De Facto

Regardless of the letter of the law, the fact remains that many writers write anonymously and academics are increasingly asked to identify them.

In courts of law, researchers with expertise in linguistics, computer science, and stylometry have acted as expert witnesses for decades now in criminal and civil disputes.

Outside of courts, academics have published or given pronouncements to journalists in most news-worthy instances involving high-profile anonymously written works, including Primary Colors (a 1996 novel satirizing the Clinton Presidential campaign), The Cuckoo's Calling (a 2013 novel revealed to be the work of J.K Rowling), the Wanda Tinasky letters (dozens of eccentric and creative letters mailed to local California newspapers from 1984-88, which academics proved were not the work of Thomas Pynchon), and many more. In all of these instances, journalists and academics have not discussed the moral or legal right to make such analysis; they have simply done it.

Ethics and stylometry

The purpose of the proposed paper is, through an analysis of different legal frameworks, to highlight the different norms and assumptions that surround the "un-masking" of anonymous authors. The radical difference in EU and US legal approaches proves that opinions can differ, and that serious debate needs to be begun among researchers. As the technology and approaches to stylometry yield increasingly accurate results, it is time for the digital humanities community to begin to discuss ethical standards.

Long papers

Representations Of Race: Mining Identity In American Fiction, 1789-1964

Mark Andrew Algee-Hewitt
mark.algee-hewitt@stanford.edu
Stanford University, United States of America

J.D. Porter
jdporter@stanford.edu
Stanford University, United States of America

Hannah Walser
walser@stanford.edu
Stanford University, United States of America

Introduction

If, as Ta-Nehisi Coates has suggested, the most lasting traumas of United States history can be traced back to Americans' "need ... to think that they are white" (2015), what might digital literary studies be able to tell us about this self-inflicted delusion? Specifically, how might quantitative textual analysis help us reconstruct the process by which ethnicities, nationalities, religious groups, and other identity categories became dominated by the all-encompassing category of race? One might expect literature to present a complex field of racial discourse, one in which what sociologist Matthew Snipp calls "administrative definitions of race" (2010) coexist with self-identifications and ethnic stereotypes, colonialist fantasies and half-forgotten family trees. Literary scholars specializing in race and ethnicity have established a rich tradition of close readings that attempt to disentangle this discourse within particular texts, generating information about how individual writers navigate race in the United States. How would this understanding change, however, if we expanded our scope to hundreds of writers and two centuries of American history? What discontinuities or consistencies might we find in the language associated with different ethnicities? Would historical changes in civil rights, immigration, and territorial expansion be visible on the level of fictional discourse?

Corpus and Methods

To approach these questions, we have assembled a corpus of 193 works of American fiction across a range of genres. Our selection begins in 1789: the year of Washington's election and the establishment of the Constitution, it also saw the publication of William Hill Brown's *The Power of Sympathy*, considered the first American novel. 1964, on the other hand, marked the beginning of President Lyndon Johnson's systematic immigration policy reforms, culminating in the Immigration and Nationality Act of 1965, which finally erased the restrictive quotas that had limited entry into the U.S. for non-white, non-western European populations. Within this extended period, major changes in government policy toward various racial and ethnic groups -- the Indian Removal Act, Emancipation, the Chinese Exclusion Act, the Immigration Act of 1917, the internment of Japanese citizens -- provide cardinal points that guide our analysis and raise fundamental issues about the relationship between fiction and the world it represents. What kinds of socio-political shifts make a difference in literary characterization? Can literature change the direction or accelerate the pace of social change, as in Abraham Lincoln's oft-quoted but probably apocryphal claim that Harriet Beecher Stowe's *Uncle Tom's Cabin* incited the Civil War?

To probe our assumptions about the language of identification in the novel, we combine methods that both investigate the formal features of the novels as a whole and extract racialized discourse as it attaches to particular characters. We suggest that the semantics of identity, whether racial, cultural, ethnic or national, operate at two discrete levels: 1) embedded within discourse such that it acts as a background to the particular worldview of the text; 2) at the level of character, where the lexemes of identity become a self-aware system of description, whether leveraged by the narrator, in the reactions of other characters or internally as part of a character's self-articulation. We argue through this project that characters embody a set of racialized identifiers that operate against a set background understanding of the meaning of identity within the text -- a dialectic between intratextual characterization and intertextual stereotyping that has its origins in, but expands significantly upon, the model of marginalized characters articulated in Alex Woloch's *The One vs. the Many* (2003). The goal of our project is to tease apart these two levels of discursive identity in order to reassemble a new history of the discourse of race in American fiction as it evolves against the backdrop of history and a changing set of aesthetic principles in novel writing.

Phase One: The Racial Unconscious

The first stage of this investigation examines the discourse of identity categories as they propagate throughout our corpus. We begin with a set list of various racial determinants that include national origins (German, Italian), ethnic identifiers (Jew, Arab) and racialized categories (Negro, Indian) and identity the pattern of language that attaches to these descriptors over time. In our first pass, we extract the collocates of each of our terms and identify which, if any, are significantly distinctive of that term.[1] For example, the term "foreign" appears significantly often as a collocate of the names of European countries, but never within the vicinity of racial descriptors, or ethnic identifiers. We then extend the process and trace a new set of second order distinctive collocates from the terms we have identified, to see which trace back to our initial set of identifiers and which introduce new discourses into the semantics of race. By visualizing our results as a dynamic network of interconnected language we trace the connections between our primary identifiers, as well as how these relationships change over time. To extend the above example, in the nineteenth century, a language of "foreignness" is connected with European nationalities, while "America/n" is distinctively used as a descriptor of African Americans. By the mid-twentieth century, the terms describing African Americans shift away from the emphasis on their "Americanness" and instead incorporate a set of terms, such as "descent" and "blood" borrowed from the foreign discourse of immigration. Such analyses can help us to identify in precise historical detail both the moment at which particular national or ethnic groups became American and the related but not the identical moment at which they, in Noel Ignatiev's phrase, "became white" (1995) -- that is, when the language surrounding those groups became unmarked. At the same time, finding consistencies in the language applied to different racial or ethnic groups at different historical moments grants support for Theodore Allen's claim that the concept of race names "a pattern of oppression (subordination, subjugation, exploitation) of one set of human beings by another," where the "phenotypical" identity of those sets is less important than the structure of their relationship (2012). This work builds upon the previous work of the members of the Literary Lab on the language of human identification in Anthropology, presented at the 2015 Digital Humanities conference and forthcoming in *Current Anthropology*, although it represents a substantial methodological extension over this early project, as well as shifting the emphasis from scholarly writing to literary representations of identity.

Phase 2: Identity as Cultural Construction

The second phase of our project examines the construction of individual characters against this backdrop. This not only allows us to observe which characters actively resist the discourse of the period in which they were created, but also how the evolution of this terminology as applied to individuals differs from the cultural construction of identity as a historically contingent socio-cultural phenomenon. That is, does a character described as "black" inherit the descriptors of identity from the language of the period or is discourse radically more individuated on a character by character basis? To test this assumption, we adopt a similar approach to the "BookNLP" developed by Bamman, Underwood and Smith (2014), using Named Entity Recognition to extract characters coupled with a set of scripts to perform co-reference resolution. We then tag the corpus for part-of-speech and extract the distinctive adjectives that appear in dependent positions within 50 words of each mention of the character. This allows us to identify, with reasonable precision, a descriptive terminology for each character. We then tag each character for the racial, ethnic or national identity that is given in the novels and compare our descriptive discourse for characters who embody similar identities across texts, using a set of semantic network diagrams that allow us to trace the contiguities of identity both across genre and across time.

Conclusion

By combining these methods, we are able to see, for the first time, not only how the distinctive language of identity alters over the history of the American novel, but how the discourse of characterization functions as both a vehicle for the standard tropes and stereotypes for identity, as well as a point of resistance to the dominant representational language of a given period. It also provides new insights into the process of characterization, especially in regard to the representation of immigrant or minority identities.

Bibliography

Algee-Hewitt, M. and Algee-Hewitt, B. (N.D.). Finding the place of race in anthropological discourse: a digital textual analysis. *Current Anthropology*, forthcoming.

Allen, T. W. (2012). *The Invention of the White Race.* London: Verso, vol. 1.

Bamman, D., Underwood, T. and Smith, N. A. (2014). A Bayesian Mixed Effects Model of Literary Character. *Proceedings of the 52nd Annual Meeting of the Association for Computation Linguistics.* Baltimore, Maryland, pp. 370-79.

Coates, T. (2015). *Between the World and Me*. New York: Random House.

Ignatiev, N. (1995). *How the Irish Became White*. New York: Routledge.

Snipp, C. M. (2010). Defining Race and Ethnicity: The Constitution, the Supreme Court, and the Census.

In *Doing Race: 21 Essays for the 21st Century.* New York: Norton.

Woloch, A. (2003). *The One vs. the Many*. Princeton: Princeton University Press.

Notes

[1] Significance is determined using a Fisher's Exact test to measure the observed values against the expected frequency of the term as a collocate, using an alpha of 0.05.

Comparing Architectural Floor Plans: New StrategiesNew Digital Tools For Architectural Historians

Patricia Alkhoven
patricia.alkhoven@meertens.knaw.nl
Meertens Institute (KNAW), Netherlands, The

Ronald Stenvert
stenvert.utr@net.hcc.nl
BBA (Bureau voor Bouwhistorie en Architectuur), Netherlands, The

Sophie Elpers
sophie.elpers@meertens.knaw.nl
Meertens Institute (KNAW), Netherlands, The

Triggered by an article by Lev Manovich *How to compare one million images?* (Manovich, 2012: 249-78) we were intrigued to find out "how to compare and analyse thousands of architectural floor plans"? The reason for this interest was the availability of the archive of the *Bureau Wederopbouw Boerderijen* (Office for Farm Reconstruction, 1940-1955) with data of 7700 farmhouses that were destroyed during World War II and were subsequently rebuilt – reconstructed or redeveloped in a new form – between 1940 and 1955 (Elpers, 2008).

Since it seemed impossible to analyse almost 8000 drawings by hand we decided to study the possibility of automatic identification and categorization of these floor plans. For this paper, we explored several strategies and digital tools to find out more about their usefulness in art and architectural history in general to answer questions about change, distribution of spaces, typology, genre, etc.

Automation of architectural floor plans

By examining the differences between and changes of the floor plans of the reconstructed farms from the period between 1940 and 1955 and by comparing them with reconstruction sketches of the old destroyed farms, we hope to be able to answer questions such as: how was the struggle about tradition and modernization which determined the debate about farm construction in the middle of the twentieth century (Elpers, 2013) resolved in practice? This question focuses on the maintenance or abandoning of the so-called 'streekeigen bouwen' (traditional regional building)(Lambert, 2007) and related planning principles. Which traditional elements were retained and which were not? Which new elements were added? Can any general structural patterns be recognized?

For this aim we would like to be able to automatically categorize and recognize elements in the architectural drawings. Following the visualization method and software (ImagePlot) by Manovich we would also like to create clusters of images of farms ordered by morphological structure, type, architect, region and timelines as output.

As architectural historians with limited programming knowledge, we started looking for existing software to find answers. Search engines usually use databases with (meta) data about objects. In our case we would rather be able to study the drawings directly on its visual content. We hoped that a combination of image processing software, techniques for pattern recognition, OCR and vectorization would bring us closer to our goals. Following information about conventions of farmhouses e.g. specific rooms, and workspaces like sheds could be labelled and found. This method, that we use in our study, has been, amongst others, described by Hansson (1998).

Preprocessing

A test-set of about 1000 images was scanned in 300 dpi TIFF: 400 reconstruction drawings and 600 new floor plans. For each farm we have two sets of scans available: drawings of the old, destroyed farms and plans of the new farms. Further, we have built a repository with the TIFF-images in relation to a simple database.

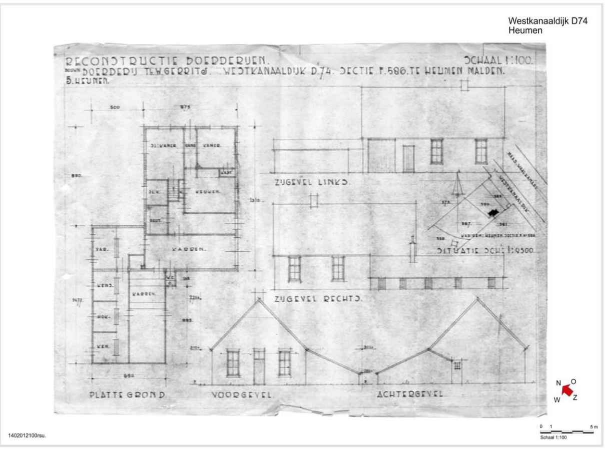


Illustration 1. Image of original architectural drawing Westkanaaldijk D74, showing plan, views and situation plan

The architectural or engineering drawings show not just one floor plan of a farm but also plans of other floors, sheds and outbuildings, cross sections and several views, a drawing of the lay-out of the premises etc. Often several details are depicted, sometimes added in a separate drawing. Although the human eye can instantly distinguish which element is the main floor plan, one has to teach the computer how to find it. In order to isolate the floor plan from all the other parts of information, the white empty spaces surrounding the different elements in a drawing could be taken as a division space, separating the individual elements. First, we had to isolate the floor plan in the drawing and deconstruct its information. Unfortunately we could not find software that could ingest the drawings and produce the isolated floor plans, without programming.

As a consequence, since it appeared too complicated to automatically find the position of the floor plan on the drawing, we skipped this part for the moment and prepared manually a new set of cut-out images of the floor plans.

With a simple image-processing program it appeared possible to manually deconstruct the contents of a drawing. In this way we could find out what steps were needed to automate the process. A drawing of Westkanaaldijk D74 in the village of Heumen (middle east of the Netherlands) was selected as a random example for the first test that shows the following manipulations:

- OCR-text
- Vectorize the floor plan
- Floor plan with text labels
- Walls (thick, medium, without openigs)
- Walls (thick, medium, with openings and numbers)

Although most techniques could be carried out with simple imaging software, such as CorelDraw – and we tried some other free online programs such as Inkscape, VectorMagic, Gimp, etc – the number of actions or manipulations just for one drawing appeared quite high. Although it works reasonably well for one drawing, this is exactly the process that we want to automate and connect with a search system.

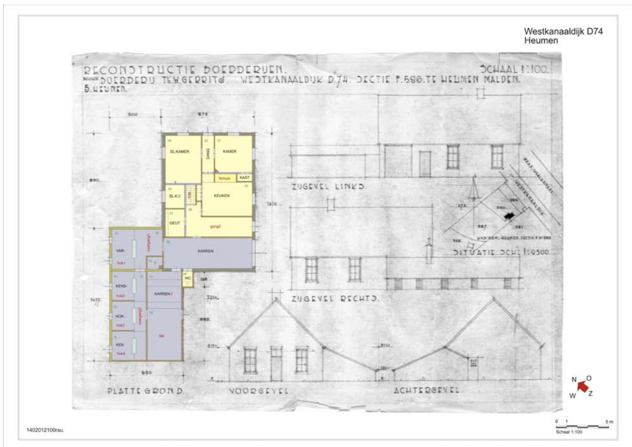


Illustration 2. Architectural drawing of Westkanaaldijk with vectorized plan and labelled spaces.

To scale up the process a bit, in a test with 10 floor plans of farms in the city of Groesbeek (middle east of the Netherlands) the following manipulations were carried out:

- the plan of ground floor and first floor were manually cut out
- images were converted from heavy tiff to the lighter jpg
- reconstruction drawings were converted from 1:200 to 1:100
- resolution of new plans were converted from 300 dpi to 150 dpi
- x/y axes were calibrated
- a central axis was determined
- it was noted that the drawings were not always on the right scale, while the inscribed measurements were noted correctly
- rooms are taken to be rectangles (which can be corrected later, but in most cases will not affect the results)
- the main spaces of farmhouses are accurate, while the interior spaces should be estimated on sight

It appeared that most walls have static measurements that can be brought into a library: the thickness of the walls depends very much on the type of bricks they are built with. These bricks have common measurements such as the *Waalformaat* (210 x 100 x 50 mm). The thickness of a wall therefore varies from 27 cm to 6 cm. The different sizes could then be coded by colour and compared.

Deconstructing drawings, isolating elements and detecting symbols

In order to deconstruct each drawing (see e.g.: Henderson, 2014; Dosch, 2000) the following steps were determined and carried out:

- Building a repository with images and database (same scale, direction)
- Finding the position of the floor plan on the drawing
- Determine (outlines of) spaces / rooms
- OCR all text information
- Recognizing walls (determining and labelling exterior and interior)
- Recognizing specific rooms (determining and labelling livingroom, kitchen, bathroom, hall, bedroom, stables etc.)

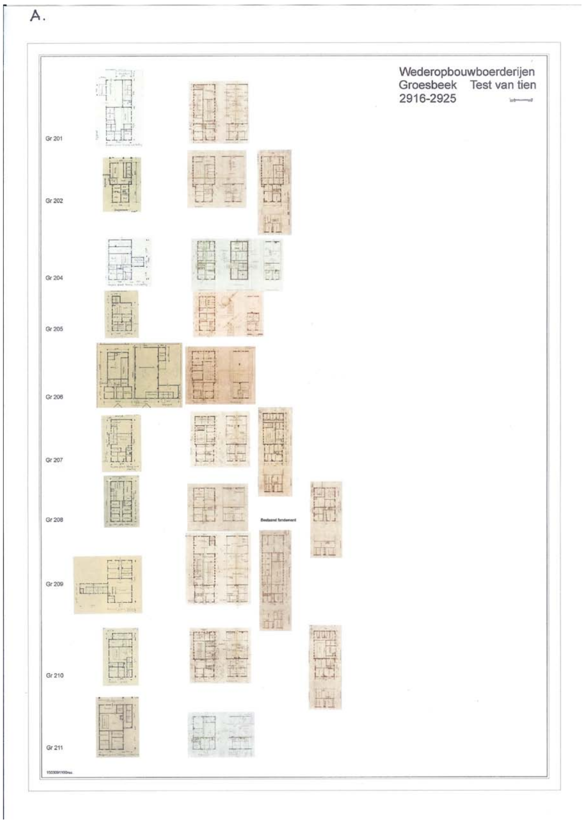


Illustration 3. Cut-out plans showing ground floor and other floors

- Determining difference between reconstruction drawing (original state) and new plan and difference between reconstruction drawings from different years

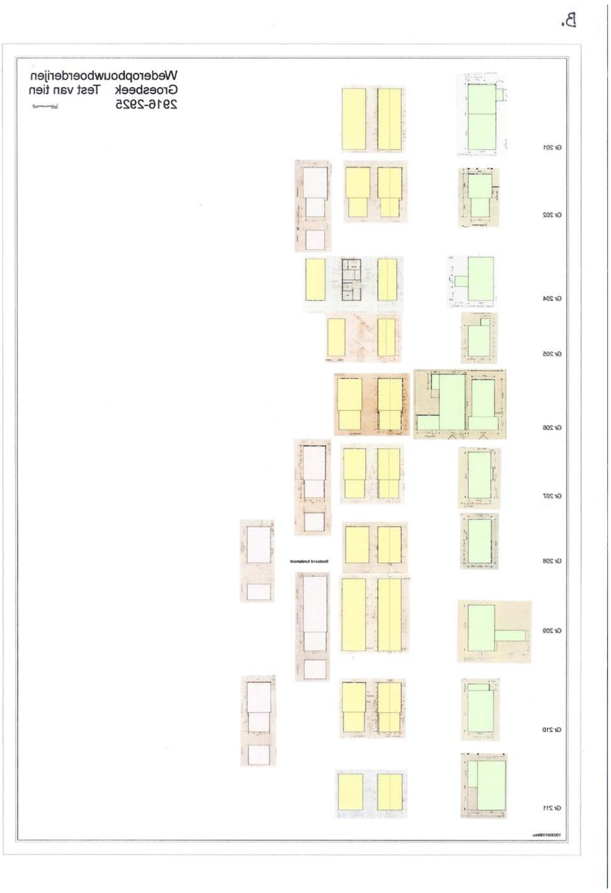


Illustration 4. Cut-out floor plans, vectorized showing the volumes of spaces.

To carry out the recognition part such as wall detection, text recognition and labelling, and making comparisons of old and new, the approach of Ahmed and Liwicki (Ahmed et al. (2011; 2014) appears promising for solving some of our problems. Although they focus on using sketches to be recognized and compared with floor plan drawings in a repository, we would like to see how their system works for us with our repository of cut-out floor plans.

Timelines can relatively easy be visualized using the metadata - with the tag "year" - in the database example (Borner, 2010; Alkhoven, 1998; 2008). Other tests with visualization software such as ImagePlot produced interesting and beautiful views of the drawings but could not yet provide the required results for analysis. More experimentation and possible adaptation of the software is needed.

Next steps

We set out to find answers on how we could compare thousands of drawings of floor plans, what digital tools and best practices are currently available to analyse them and how we could visualize differences in sets of floor plans. In this first stage of the process we have gathered information about the preprocessing stages and we were able to perform some experiments with the scanned architectural drawings.

We discovered many – free–tools but it turned out difficult to distinguish among them and to choose the best ones that could solve each part of the problem. We have tested several apps but since it was not our objective to deal with coding, we could not adapt the software and we were therefore limited in doing our research. As a consequence much had to be done manually in a computer-assisted way.

Since our overall problem appeared a bit too ambitious to carry out within the given boundaries of time and budget, we have limited ourselves to producing a list of ingredients and a well-argumented recipe formulating the next sequence of steps to take. These form the basis for a new research proposal for the Digital Humanities' call in The Netherlands. A long article with more detailed information about the process, experiments, and results, will be ready later this year.

Bibliography

Ahmed, S. et al. (2011). Improved automatic analysis of architectural floor plans. *Document Analysis and Recognition (ICDAR)*, pp. 864-68.

Ahmed, S. et al. (2014). Automatic Analysis and Sketch-based Retrieval of Architectural Floor Plans, *Pattern Recognition Letter*, 35, pp. 91-100.

Alkhoven, P. (1993). *The Changing Image of the City. A Study of the Transformation of the Townscape by means of Computer-Aided-Architectural-Design and Visualization Techniques. A case study: Heusden.*

Alphen aan den Rijn.

Alkhoven, P. (1998). Computer Visualisation as a tool in Architectural Historical Research. *Architectural and Urban Simulation Techniques in Research and Education*, Delft University Press, pp. 16-23.

Berry, D. (2012). *Understanding Digital Humanities*, Palgrave: Macmillan.

Börner, K. (2012). *Atlas of Science. Visualizing what we know.*

Dosch, Ph. et al. (2000). A complete system for the analysis of architectural drawings. *International Journal on Document Analysis and Recognition*, vol.**3**: 102-16.

Elpers, S. (2008). Het archief van het Bureau Wederopbouw Boerderijen, *Vitruvius.* **1**(3): 40-47.

Elpers, S. (2013). *Erfenis van het verlies. De strijd om de wederopbouw van boerderijen tijdens en na de Tweede Wereldoorlog* (unpublished PhD thesis) Amsterdam: University of Amsterdam).

Hanson, J. (1998). *Decoding Homes and Houses,* Cambridge.

Henderson, T.C. (2014). *Analysis of Engineering Drawings and Raster Map Images.*

Lamberts, B. (2007). *Boerderijen. Categoriaal onderzoek wederopbouw 1940-1965*. Zeist.

Manovich, L. (2012). How to compare one million images? *Understanding Digital Humanities*, pp. 249-78.

Geocoding Thousands of Fiscal Records: Methodological Approach for a Study on Urban Retail Trade in the Belle Époque

Daniel Alves
alves.r.daniel@gmail.com
IHC, FCSH, New University of Lisbon, Portugal

In August 1902 the British newspaper The Times inserted an article announcing "The passing of the grocer" as a result of a crisis in the small retail trade that apparently had bankrupted more than 900 stores (Winstanley, 1983). Across the Atlantic, Canada's shopkeepers faced a crisis equally significant, with overabundance of shops and a lot of bankruptcies (Monod, 1996). The same happened in continental European cities like Paris or Milan, in the final decades of the nineteenth century (Nord, 1986; Morris, 1993). Much of what happened this time was the result of an demographic growth in the cities in the nineteenth century that almost everywhere would have important consequences in the reorganization of urban space, leading to changes in their economic and social geography. Lisbon was no exception and the city's retail trade, after a phase of growth in the 1880s, faced a crisis in the last ten years of the century (Alves 2012). This occurred while the city was transformed, witnessing an expansion into new urban areas to accommodate an increasing volume of inhabitants.

Through the analysis of the geographical distribution of Lisbon's retail trade, between 1890 and 1910, it is possible to verify the impact of this crisis on the way small businessmen apprehended the urban space, as well as the opportunities and risks that those changes could represent to them. The study is based on a very detailed information about the localization and characteristics of every single shop in the city streets, gathered from fiscal sources in the municipal archives (around sixteenth thousand records for each of the three years, 1890, 1900 and 1910) and analyzed through the use of spatial statistical analysis tools. This volume of information and the potential introduced by a Digital Humanities / Spatial Humanities approach, brings in methodological challenges regarding the digitization and geocoding process of several thousands of records, as well as opens new research possibilities and new research questions, namely: about the role of women in the retail trade business, since previous work based on smaller sets of data analyzed with traditional qualitative methods had misrepresented or even ignored gender issues in the retail business (Alves, 2012); or about the influence of the dwelling rents in determining changes in the social space of a city in the Belle Époque, only possible by the digital analysis of a very large, temporal and geographical encompassing amount of data for all the city.

In terms of the methodology two were the challenges we faced in two distinct stages of research, data collection and its georeferencing. The first challenge relates to the difficulty in scanning a volume of data of this nature, only available on archive, with no available funding to carry out a project of mass digitization of the original tax records. This was associated with the fact that all sources are handwritten, drawn up by several employees, with equally different handwritings, which would make it impossible, given the current state of handwritten character recognition technology (Beatty, 2010; Brumfield, 2014), for an automatic or even semiautomatic data treatment. The solution followed a close approach to crowdsourcing projects (Causer et al., 2012), using shared databases and collaborative work, a method already developed with excelent results in previous studies (Alves and Queiroz, 2013; Alves and Queiroz, 2015). The second challenge went through georeference about fifty thousand addresses obtained in this documentation, using essentially the computer's processing power. It is recognized that there are advantages and disadvantages in using several geocoding methods (Zandbergen, 2008). Its application to Lisbon is made difficult due to the fact that the city have gone through a deep urban morphology transformation and experienced significant changes in its street names throughout the twentieth century (Alves, 2005; Oliveira and Pinho, 2006). Nevertheless, it remains one of the best methods for automatic assignment of geographic coordinates. The challenge was to think of a process that could overcome the difficulties listed, without having to go through the full reconstitution of the urban network. Not least because the available sources do not made possible the recreation of all the buildings and their functional classification in a GIS environment, as was achieved in other projects (Dunae et al., 2013). The option went through reconstruction of the existing streets network of the time, based on geo-referenced digital cartography. Thus an addresses' database was created, with slight adaptations on the geocoding algorithm of the GIS platform used, allowed for a success rate of around 90%.

As for georeferencing, the first and greatest of all the problems is related to the urban changes that the city has undergone over the last 120 years. On the one hand, strong population growth led to the expansion or changes in the corresponding urban area, even If we take into account just the three years analyzed. The city of 1890 is very different from the one in 1910, in the layout of the streets, in the expansion into new areas, construction or renovation of its buildings. On the other hand, Lisbon had the particularity to overcome, between 1890 and the present, four very different political regimes that have left a peculiar and deeply transformative mark in the city's toponomy. Even if the urban area was stable, only the change of street names over more than a century of profound political changes, would pose great challenges to an automated geocoding process.

For this there was first the need to rebuild the streets network of the time, trying to recreate the map of Lisbon of the Belle Epoque. This map was fixed for 1890 and then it was possible to apply the normal geocoding techniques, adapting either the collected data source, or the software algorithm used, in order to overcome difficulties so small, but so significant in the final results, such as the fact that the software was incapable to deal with some Portuguese names and characters or to recognize certain types of streets that were used at the time but ignored or underused today. Or the fact that the names of some streets are identical or very similar in different areas of the city. This is an iterative process of trial and error so as to refine the best model data that maximizes the results obtained.

However, there was still the issue of street names changing over time. For the period in question this problem is not very complicated, since only at the end of 1910 took place the first regime change, the Republican Revolution, with the consequent wave of place names changed. But there was almost annually, specific changes in street names. These changes were incorporated in the database, maintaining the address structure already georeferenced for 1890, allowing to incorporate the data for the same streets that appeared in the sources with different names.

These two ways to overcome the problems with georeferencing were only possible due to the existence of sources, either cartography or streets itineraries, which allowed to go adapting the original map so as to incorporate, with a similar success rate, the 1900 and 1910 data.

As for the quality or accuracy of georeferencing, it is obvious that the ideal would be to have the possibility to rebuild not only the network of city streets, but also the actual location of the various housing/trade blocs. Unfortunately no sources for this level of detail are available. In the original shapefile map, which represented the actual city streets, in 2012, the streets were targeted with the actually existing building numbers in each block. Since we have no way of saying, at this stage of the research, how many buildings existed on every block in 1890, it was decided to distribute all points along the entire length of the street, according to the geocoding algorithm. Obviously, this option can cause some significant deviation in less consolidated urban areas of the city, but it is also true that the areas where the overwhelming majority of small businesses traditionally where implemented account for already established streets. So that problem here is much lower, and the geocoding accuracy is much higher.

I did not use crowdsourcing primarily due to the available time for the collection of data and to poor accessibility to the original sources. Seems odd to state this because apparently these are precisely justifications for using crowdsourcing. However, it is known, also through the literature already cited in the abstract, that the use of crowdsourcing can be a lengthy process and not always easy to check on the quality of final results. Furthermore, it requires easier access to sources through digital copies because not all potential participants have availability or even competence for archive work. In our case, the sources correspond to thick volumes of bound sheets, deposited in a municipal archive without major conditions for local consultation and without the financial ability to carry out its full or partial scan. There was also the problem of the quality of the data, given that we are talking about handwritten information sometimes difficult to read. In this sense, the use of a collaborative work, through volunteer history students with availability and sensitivity to archival work was decisive. This data collection was made via a PostgreSQL database, shared with all students through ODBC, so that all data collected by one of the students become available for subsequent use by others. This process, already tested in other project with very good results (Alves and Queiroz, 2013; Alves and Queiroz, 2015) , saved up a lot of time in data collection and redundancies were avoided, because a new address o r a new commercial activity detected by one of the students in the sources once registered in the database, was automatically available and could be used/selected by all the others. Even if the original record contained an error, since it was registered only once in the database, because the data model prevented duplicates, the validation task was also facilitated.

The reconstruction of the 1890 street map, base of all subsequent work, was made through a current map in shapefile format, provided by the Municipality. This map was superimposed on a digital copy of an old city map properly georeferenced. The mere overlap of the two maps identified a broad set of streets that did not exist at the time (eliminated from the shapefile) and others that changed their layout (fixed in the shapefile). Using streets itineraries, namely one for 1890, very complete, mentioning the name, location and building numbers of all city's streets, it was possible to correct the shapefile street names that had gone through changes over these 120 years. The link between this corrected shapefile and the shops' addresses collected in the shared database allowed for geocoding, with minor adjustments to the software algorithm.

Overcoming these issues, it was possible to get a far more dynamic source, more accessible and manageable, able to respond quickly to old research questions or to introduce new perspectives about Lisbon and its retail trade in the Belle Époque. Just to mention one possibility, in the Lisbon municipal archive there are many available data that could be crossed with the information collected on shops and shopkeepers. In particular, it has data on electoral census and elections, with voters lists organized by parishes and addresses, for several years of the nineteenth and twentieth century and this data can also be mapped and analyzed using the methodology now developed. The same can happen with data on primary education, for example, because this fund has information on the addresses of the students also. Given that the addresses of the shops are being collected in a separate table and that it allows to be connected to either the GIS or to other tables with different data, whether on small businesses, on elections or on education, the use of this methodology will allow for results in other projects or to add new information relevant to the history of the city without the need to start everything from scratch.

In this sense, while recognizing the limitations of GIS (Boonstra, 2009; Bodenhamer et al., 2010; Silveira, 2014), it must be highlighted the capabilities of its spatial analysis, processing and data visualization tools, that has been particular useful in the context of urban history (Hillier, 2010; DeBats and Gregory, 2011). In recent years, the use of GIS for the study of cities and their retail trade (Beascoechea Gangoiti, 2003; Mirás Araujo, 2008; Bassols and Oyón Bañales, 2010; Novak and Gilliland, 2011; Ünlü, 2012) has enhanced the theoretical framework and the possibility of international comparative studies, which represented a definitive stimulus to the application of these methodologies to this case study. To this matter it is useful to mention that a session comparing four case studies for Iberian port cities will occur in the ESSHC in Valencia, this year, putting together comparative analysis about Barcelona, Bilbao, Coruña and Lisbon, in the first decades of the twentieth century (https://esshc.socialhistory.org/esshc-user/programme?day=55&time=145&session=3003).

Bibliography

Alves, D., (2012). *A República atrás do balcão: os Lojistas de Lisboa e o fim da Monarquia (1870-1910)*, Chamusca: Edições Cosmos.

Alves, D. (2005). Using a GIS to reconstruct the nineteenth century Lisbon parishes. *Humanities, Computers and Cultural Heritage. Proceedings of the XVIth international conference of the Association for History and Computing*. Amsterdam: Royal Netherlands Academy of Arts and Sciences, pp. 12–17.

Alves, D., Queiroz, A. I. (2015). Exploring Literary Landscapes: From Texts to Spatiotemporal Analysis through Collaborative Work and GIS. *International Journal of Humanities and Arts Computing*, **9**(1): 57–73.

Alves, D., Queiroz, A. I. (2013). Studying urban space and literary representations using GIS: Lisbon, Portugal, 1852-2009. *Social Science History*, **37**(4): 457–81.

Bassols, M. G. and Oyón Bañales, J. L. (2010). Retailing and proximity in a liveable city: the case of Barcelona public markets system. In *REAL CORP 2010: CITIES FOR EVERYONE. Liveable, Healthy, Prosperous*. Vienna, pp. 619–28.

Beascoechea Gangoiti, J. M. (2003). Jerarquización social del espacio urbano en el Bilbao de la industrialización. *Scripta Nova*, **VII**(142): 1–19.

Beatty, J. (2010). Historical Documents in a Digital Library: OCR, Metadata, and Crowdsourcing. *Lemonade and Information*. Available at: https://lemonadeandinformation.wordpress.com/2010/05/28/historical-documents-in-a-digital-library-ocr-metadata-and-crowdsourcing/ (accessed October 25, 2015).

Bodenhamer, D. J., Corrigan, J. and Harris, T. M. (Eds.) (2010). *The spatial humanities: GIS and the future of humanities scholarship*, Bloomington: Indiana University Press.

Boonstra, O. (2009). Barriers between historical GIS and historical scholarship. *International Journal of Humanities and Arts Computing*, **3**(1-2): 3–7.

Brumfield, B. W. (2014). Collaborative Digitization at ALA 2014. *Collaborative Manuscript Transcription*. Available at: http://manuscripttranscription.blogspot.pt/2014/07/collaborative-digitization-at-ala-2014.html (Accessed October 25, 2015).

Causer, T., Tonra, J. and Wallace, V. (2012). Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham. *Literary and Linguistic Computing*, **27**(2): 119–37.

DeBats, D. A. and Gregory, I. N. (2011). Introduction to Historical GIS and the Study of Urban History. *Social Science History*, **35**(4): 455–63.

Dunae, P. A. et al. (2013). Dwelling Places and Social Spaces: Revealing the Environments of Urban Workers in Victoria Using Historical GIS. *Labour/Le Travail*, 72, pp. 37–73.

Hillier, A. (2010). Invitation to Mapping: How GIS Can Facilitate New Discoveries in Urban and Planning History. *Journal of Planning History*, **9**(2): 122–34.

Mirás Araujo, J. (2008). The Commercial Sector in an Early-Twentieth Century Spanish City, La Coruña 1914-1935. *Journal of Urban History*, **34**(3): 458–83.

Monod, D. (1996). *Store wars: shopkeepers and the culture of mass marketing, 1890-1939*, Toronto: University of Toronto Press.

Morris, J. (1993). *The political economy of shopkeeping in Milan, 1886-1922*, Cambridge: Cambridge University Press.

Nord, P. G. (1986). *Paris shopkeepers and the politics of resentment*, Princeton: Princeton University Press.

Novak, M. J. and Gilliland, J. A. (2011). Trading Places: A Historical Geography of Retailing in London, Canada. *Social Science History*, **35**(4): 543–70.

Oliveira, V. and Pinho, P. (2006). Study of urban form in Portugal: a comparative analysis of the cities of Lisbon and Oporto. *Urban Design International*, **11**(3): 187–201.

Silveira, L. E. da (2014). Geographic information systems and historical research: an appraisal. *International Journal of Humanities and Arts Computing*, **8**(1): 28–45.

Ünlü, T. (2012). Commercial development and morphological change in Mersin from the late nineteenth century to the mid-twenties: modernization of a mercantile port of exchange in the Eastern Mediterranean. *Planning Perspectives*, **27**(1): 81–102.

Winstanley, M. J. (1983). *The shopkeeper's world, 1830-1914*, Manchester: Manchester University Press.

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, **32**(3): 214–32.

Unlocking The Mayan Script With Unicode

Deborah Anderson
dwanders@sonic.net
UC Berkeley, United States of America

Carlos Pallán Gayol
pallan.carlos@gmail.com
University of Bonn, Germany

Problem

At present, three hieroglyphic scripts are encoded in the Unicode Standard: Egyptian, Anatolian, and Meroitic. The historic Mayan hieroglyphic script is not yet in Unicode, in part because its complex clustering poses special encoding problems.

Besides the difficulties inherent in the Mayan script, writing a Unicode proposal for Mayan involves additional challenges, requiring significant time commitment and travel. Early meetings between with Unicode experts are a desideratum in order to identify the best technical approaches to handle Mayan text issues. Then the proposal needs to be written – complete with all the required details for the entire character repertoire. When the proposal goes before the two standards committees, the authors need respond to any questions, and make revisions to the proposal as needed. In addition, proposers need to commit to seeing the process through to the end, which can take between two to five years, or longer.

In short, the twin challenges of representing and displaying Mayan clusters on computers and writing a Unicode proposal (especially for anyone new to the process) present a huge hurdle.

Background on Mayan clustering

Mayan signs appear in clustered glyph-blocks or "collocations," which could be modified by the masterful Mayan scribes with advanced "tools," such as ligatures, conflation, infixation, superimposition, pars pro toto and full-figure variants (see Fig s. 1 and 2). The visual complexity of Mayan requires mechanisms that go beyond standard script encoding approaches.

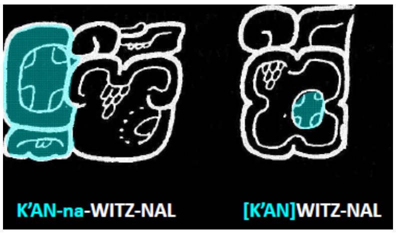


Fig. 1 Infixation of the highlighted logogram K'AN (shown on left, infixed as a circle on the right)

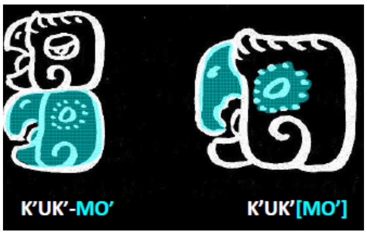


Fig. 2 Conflation of the highlighted logogram MO' (shown on the left, conflated on the right)

Solution

A recent multidisciplinary collaboration established between UC Berkeley's Script Encoding Initiative (SEI) and the MAAYA Project aims to employ new methods combining linguistics, Mayan epigraphy, digital palaeography and computer vision to overcome some of the major challenges preventing the encoding of Mayan hieroglyphs in the Unicode Standard. Some of the strategies envisioned rely on already existing MAAYA-Project resources. Such resources include datasets with annotated database records for all individual glyphs and possible sign-combinations attested in Mayan hieroglyphic books and an advanced concordance functionality, that cross-references Mayan signs across existing glyph-catalogues (Gatica-Perez et al., 2014 and Hu et al., 2015).

In addition, SEI's experience in overseeing the encoding over 70 of the world script systems means Mayan experts will have a front-line guide to help the proposal through the entire encoding process – providing Unicode expertise, assisting on the authoring and review of a proposal, and presenting the proposal at standards meetings. Having direct involvement of Unicode specialists means the proposal can draw on Unicode experience to identify successful methods that can adapted and expanded to account for the extraordinary variability of Mayan signs clustered in glyph blocks.

Specific encoding issues: Egyptian format characters

In Unicode 8.0, the default display for the three encoded hieroglyphic scripts (Egyptian, Anatolian and Meroitic Hieroglyphs) is a linear listing of the characters, as shown by the example in Fig. 3.

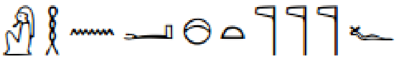


Fig. 3 Current default display of Egyptian hieroglyphs in Unicode (Richmond, 2015)

According to the Unicode Standard, display of the characters in a non-linear manner (i.e., in clusters) should be handled by a higher-level protocol (Unicode Consortium

2015: 430-31, 437), and is outside the scope of Unicode. However, clustering is more typical of the layout for most hieroglyphic scripts.

Although Egyptian hieroglyphs were encoded in Unicode in 2009, Egyptologists are currently prevented from interchanging data easily, because they have had to rely on non-standard software to handle character grouping. According to Richmond 2015, hieroglyphs need to be capable of being grouped together in plain-text - without proprietary software - in order to be truly useful.

In 2015, a new proposal was put forward for three format characters that allows basic clustering (Richmond, 2015, see Fig. 4). The characters indicate placement of a character, either above another character or alongside it, or, for the third character, identify it as forming a ligature with the following character. The three characters are based upon the conventions of the Manual de Codage, which uses ASCII characters to indicate the placement of the hieroglyphs. (Note: Since the mirroring of glyphs is handled by line direction, as specified by markup or bidirectional characters, no format characters are not proposed for that.)

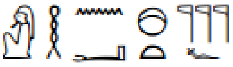


Fig. 4 Example showing the same characters as in Fig. 3, but with expected clustering, made possible with the format characters (Richmond, 2015)

The three Egyptian format characters were approved by one of the two standards committees in January 2016 (Unicode Consortium, 2016). The characters have been tested and shown to display as expected in the Universal Shaping Engine, a rendering engine in Windows 10 and recent versions of Android, Chrome, LibreOffice and Firefox.

If the characters successfully complete the approval process, simple clustering in Egyptian hieroglyphs should be capable of plain-text representation in the future, meaning that scholars would not have to rely on an ad hoc, non-standard solution. At this point, Egyptian hieroglyphs appear to only require the three format control characters for clustering, and need no other mechanisms.

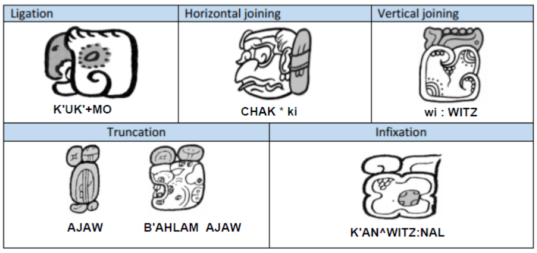


Fig. 5 Expected clustering of Mayan hieroglyphs (adapted from Richmond and Glass, 2016)

Evidence from the MAAYA project suggests that Mayan will require three script-specific format characters that serve the same functions as Egyptian, and at least two additional characters: one for truncation and another for infixation (see Fig. 5).

Hence, it appears that Egyptian and Mayan can share a common model, using format characters for clustering (with at least two additional ones needed for Mayan). In addition, advances in text display of Egyptian hieroglyphs, which have been tested successfully on the Universal Shaping Engine, may well be applicable to Mayan.

Specific encoding issues: Ideographic Description Characters

Unicode contains a mechanism to describe Chinese-Japanese-Korean (CJK) ideographs, called Ideographic Description Characters (IDCs). These characters are used to describe the layout of CJK characters, but is not used for rendering. The IDCs have been defined as capable of being extended to other scripts (Unicode Consortium 2015, pp. 679-80). (See Fig. 6.)

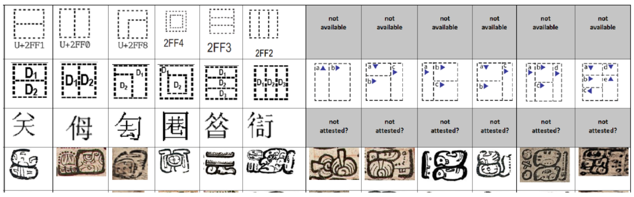


Fig. 6 Examples of Ideographic Description Characters (extended to Mayan hieroglyphs). Top row: Unicode Ideographic Description Characters (IDC); second row: Chinese IDCs (left) and Mayan descriptors (right); third row: Chinese examples; fourth row: Mayan script examples

In April 2015, Unicode experts met with the co-author, Mayan expert Carlos Pallán Gayol, and recommended the MAAYA project identify the structural patterns of Mayan characters, and use Ideographic Description Characters (IDCs) to describe them. Such a mechanism will help in defining Mayan characters (in Unicode terms), and help "unlock" the script.

Next steps

At a meeting in January 2016 with the co-author Pallán Gayol, unicode experts noted progress on the encoding model issues and suggested the MAAYA project continue to define the base set of characters, identify the structural patterns of IDCs, and verify the number of format characters needed.

Expected results

Encoding the Mayan hieroglyphs in the Unicode format will allow creation of vast open-access Mayan hieroglyphic text repositories and libraries, upon which advanced search and query functionalities relying i.e. on Optical Character

Recognition (OCR) and text-mining could be applied. (Once a script is in Unicode, an OCR engine can be trained to read the script, though the text needs to be consistent).

Thus, we argue that the ability to render any Mayan hieroglyphic text in an encoded digital format could impact on the overall accessibility, reproduction, visualization and long-term preservation of the sum of ancient knowledge recorded by the Mayan scribes on thousands of texts and inscriptions produced between ca. 250 B.C. and 1520 AD in Central America. It could also act as a model for the encoding of other hieroglyphic scripts of the Americas, including Aztec.

Funding

This work was supported by the National Endowment for the Humanities [PR-50205-15] and by the DFG (Deutsche Forschungsgemeinschaft, German Scientific Association).

Bibliography

Gatica-Perez, D. C., Pallán Gayol, S., Marchand-Maillet, J.-M. et al. (2014). The MAAYA Project: Multimedia Analysis and Access for Documentation and Decipherment of Mayan Epigraphy. In *Proc. Digital Humanities Conference* (No. EPFL-CONF-202571). http://publications.idiap.ch/downloads/papers/2014/Gatica-Perez_DH_2014.pdf (accessed 2 March 2014).

Hu, R., Can, C., Pallán Gayol, G. et al. (2015). Multimedia Analysis and Access of Ancient Mayan Epigraphy. *Signal Processing Magazine, IEEE*, **32**(4): 75-84. http://publications.idiap.ch/index.php/publications/show/3093 (accessed 2 March 2014).

Richmond, B. (2015). Proposal to encode three control characters for Egyptian Hieroglyphs. http://www.unicode.org/L2/L2015/15123r-egyptian.pdf (accessed 2 March 2014).

Richmond, B. and Glass A. (2016). Proposal to encode three control characters for Egyptian Hieroglyphs. http://www.unicode.org/L2/L2016/16018r-three-for-egyptian.pdf (Accessed 2 March 2014).

Unicode Consortium. (2015). *The Unicode Standard, Version 8.0.0.*. Mountain View, CA: The Unicode Consortium. http://www.unicode.org/versions/Unicode8.0.0/ (accessed 2 March 2014).

Unicode Consortium. (2016). Draft Minutes of UTC Meeting 146. San Jose, CA. http://www.unicode.org/L2/L2016/16004.htm (accessed 2 March 2014).

Developing Competencies in Digital Scholarship Among Humanities Scholars

Smiljana Antonijevic Ubois
smiljana@smiljana.org
Penn State University, United States of America

Introduction

This paper draws on results of a three-year ethnographic study funded by the Andrew W. Mellon Foundation and the Royal Netherlands Academy of Arts and Sciences, which explored research practices, challenges, and directions in contemporary digital humanities (DH). Within this broader set of questions and results, the current paper focuses on findings related to humanists' strategies for developing competencies in digital scholarship. The study was conducted from 2010 to 2013 at twenty-three educational, research and funding institutions in the United States and Europe, and it included case studies, surveys, in-depth interviews, and observations. The study involved 258 participants, including researchers, faculty, students, university administrators, librarians, software developers, policy makers, and funders. For more information about the methodological design and results of the study see *Amongst Digital Humanists: An Ethnographic Study of Digital Knowledge Production* (Antonijević, 2015).

Results

The majority of humanists consulted in this study reported awareness of methodological and epistemological benefits of digital research tools and methods, but also a lack of opportunity to acquire skills and knowledge that would enable them to reach beyond the "search and access" level of digital scholarship. As one assistant professor of art history put it, "I haven't used technology in my research in a pervasive way to really, really think about epistemological issues. I'm not opposed to using technology to analyze, but I haven't had a chance to learn it."

Humanists commonly identify lack of time as one of the main impediments to developing digital research skills. One aspect of this is the learning curve, and a perception that the time needed to learn new tools and methods slows down their established research process. Another root of researchers' lack of time for developing computational skills stems from the structure of disciplinary incentives and rewards in humanities disciplines. Respondents underscored that when training sessions on digital research tools and methods were organized at their universities, hardly any of the tenure-track faculty attended them: "We are not rewarded for doing that. What we are rewarded for is publishing, and going to one of those sessions takes away

[time] from our publishing. So, there's a lot of resistance" explained an assistant professor of linguistics.

With digital skills still having an unrecognized status in their disciplines and departments, interviewees said the only organized educational initiatives at their disposal were training sessions at university libraries, to which they had mixed reactions. While some found the library sessions "eye-opening" the majority did not regard them as helpful. Commonly, respondents pointed that the librarians focused on digital tools and resources, while field-specific research questions exceeded their scope. The respondents held that only their peers understand epistemological and methodological complexity of particular research problems, so they considered working with colleagues as a more effective way to develop digital skills than attending library workshops or instruction sessions. In the same way, respondents identified as the preferred type of instruction the one that does not profess to teaching about digital technologies, but about a specific humanities subject area, introducing digital tools and methods along the way. Attending library and similar workshops was thus seen as less effective then attending academic conferences where peers present results achieved through digital methods and tools. Learning by example inspires humanists to discover new tools and methods, and to apply them in their own work. As a professor of Romance languages and literature put it, "in an ideal world, I would like to see humanists teaching other humanists how to conduct research using the enhancements of digital tools; I guess that's as opposed to we bring the IT people in to teach us. I think that it should happen the way that we teach other things in the humanities, through collaboration with one another."

Dissatisfaction with existing educational initiatives might be the reason that the majority of humanities scholars consulted in this study reported not having any formal technology-related instruction, which is consistent with Siemens' (2013) findings. Instead, the respondents reported that they predominantly relay on informal channels, such as word of mouth, to learn about digital research tools and methods: "I do everything on my own, I ask around. It feels serendipitous, I sort of bump into it, or I hear a friend talk about it, or a colleague will shoot me an e-mail. It's not organized or strategic at all" related an associate professor of philosophy.

This informal learning path is linked to immediate and specific research problems scholars are facing, which makes it preferred over workshops and similar efforts where learning is often decontextualized from practice. This method also successfully makes use of one of the scholars' most scarce recourses—their time. It enables them to direct learning efforts towards tools, methods, and subjects of particular interest to them.

Informal learning also often takes place through engagement with students. The respondents explained that teaching prompts them to expand knowledge of digital tools and methods, and students challenge them to be up-to-date with technology. Even those respondents who described themselves as "technological dinosaurs" identified the need to take up digital technologies in the classroom.

In addition to class interaction, more formal initiatives where students teach faculty take place in digital humanities centers, where graduate students often work as tutors. Students consulted in this study believed that this reversal of instructional roles facilitated their understanding of the didactic principles, motivating them to develop their own pedagogic strategies:

"You're doing work with people who are the smartest people on the earth in their particular discipline. On average, they don't like to admit that they don't know something. So, it's not trying to force things upon them, but it's trying to present things in the same way that you would pedagogically teach a difficult concept. If I can present it it in such a way that leads them to discovering on their own, they have a sense of ownership of the idea, so that's "teaching the old dog new tricks" if I can say that." [A graduate student]

Conclusion

Findings of this study indicate that library initiatives and units should not remain the main locus of digital scholarship in the humanities. Instead, digital scholarship needs to be part of humanities departments and wider university initiatives, since "digital humanities will ultimately matter, or not at all, *inside* the department" (Liu, 2009: 21; italics in the original). As long as humanists' interaction with digital tools, methods, and resources is treated merely as a technical skill that can be taught by non-expert personnel, it will be difficult to achieve more substantial transformations and to motivate academics about digital scholarship. As Raley (2014) points out, "academic service staff providing skills-based training [...] and performing service work for "clueless arts and humanities scholars" can tell us something about both the field and the university" (p. 7).

Instead of skilled-based training, humanities education in digital scholarship needs a comprehensive framework encompassing epistemological, methodological, technical, and sociocultural aspects of digital knowledge production. These include developing understanding of digital and other types of data, fostering critical reflection on digital objects of inquiry, comprehending the influence of algorithmic processes on humanities investigations, and so on. Similarly, digital methods training should include systematic deliberation on methodological decisions influencing research process and results, epistemological and ethical challenges of digital scholarship, as well as making choice of digital tools and methods as best suited for specific research questions.

This study reveals that humanists favor and best learn in practice, when instruction is closely related to their area of study and when it unfolds organically, through collaboration with colleagues and students. Therefore, initiatives for developing competencies in digital scholarship among senior scholars should use a variety of collaborative learning strategies. Furthermore, these educational activities should not restrain the generative potential of digital scholarship in the humanities through the exclusive focus on research themes, methods, and skills recognized in the DH field. Humanities scholars do not necessarily need or want to be digital humanists; they do, however, need and overwhelmingly want to be scholars competent at teaching and conducting research in the digital age. As a program officer in a major humanities foundation consulted in this study pointed out, the goal should be "to fund training for scholars even if they don't want to be a digital humanist in the sense that they're building their own tools and their programming, but more along the lines of users of digital technology within their own research." This kind of funders' support to digital scholarship in the humanities is vital. Equally vital is the need that education in digital scholarship becomes administratively recognized as part of scholars' professional development included in their paid time and activity, as well as in their promotion dossiers.

Bibliography

Antonijević, S. (2015). *Amongst Digital Humanists: An Ethnographic Study of Digital Knowledge Production*. New York: Palgrave Macmillan.

Liu, A. (2009). Digital Humanities and Academic Change, *English Language and Notes*, **47**: 17-35.

Raley, R. (2014). Digital Humanities for the Next Five Minutes, *differences*, **25**(1): 26-45.

Siemens, L. (2013). Developing Academic Capacity in Digital Humanities: Thoughts from the Canadian Community, *Digital Humanities Quarterly*, **7**(1). http://www.digitalhumanities.org/dhq/vol/7/1/000114/000114.html (accessed August 6, 2014).

Prototypes as Thinking through Making. Decision Points and Evaluation in Prototyping a Visualisation Framework for Historical Documents

Florentina Armaselu
florentina.armaselu@cvce.eu
Centre Virtuel de la Connaissance sur l'Europe (CVCE), Luxembourg

Roberto Rosselli Del Turco
roberto.rossellidelturco@fileli.unipi.it
Università di Torino

Catherine Jones
catherine.jones@cvce.eu
Centre Virtuel de la Connaissance sur l'Europe (CVCE), Luxembourg

Lars Wieneke
lars.wieneke@cvce.eu
Centre Virtuel de la Connaissance sur l'Europe (CVCE), Luxembourg

Chiara Alzetta
chiara.alzetta@gmail.com
Università di Pisa

Chiara Di Pietro
dipi.chiara@gmail.com
Università di Pisa

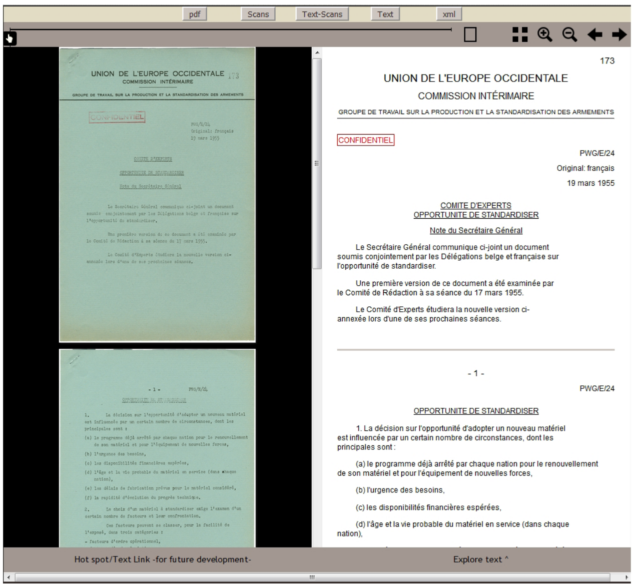
Introduction

Starting from Manovich's (2007) statement that "a prototype is a theory", Galey and Ruecker (2010: 406-07) argue that design can become "a process of critical enquiry itself", a "thinking through making" pursuit allowing the combination of digital prototyping with critical analysis. Furthermore, Pierazzo (2011: 466) brings into discussion the theoretical assumptions that may underpin the decision making process in building an edition as an "interpretative scholarly product" based on the "selection of features transcribed from a specific primary source".

The present proposal focuses on the construction of a visualisation framework allowing transformation and visualisation in the browser of XML-TEI encoded documents on European integration history. The tool developed for this purpose, the Transviewer, uses a combination of XML, HTML, XSLT, CSS and JavaScript technologies. The addressed research questions are related to the analysis of this prototyping case, viewed as a dynamic and iterative process of evaluation, adjustment, decision making, adaptation and in-house development. Our standpoint, inspired by the above mentioned studies, is that such type of analysis can shed light on the theoretical and practical questions, at the crossroad of tradition and new ground, involved in the creation of scholarly digital tools.

The Transviewer prototype

The Transviewer concept consists of building a framework for the publication of European history documents on the CVCE's Website, from treaties, official declarations and meeting reports to letters and interview transcriptions. In a first phase, a pilot testing set for the prototype (Figure 1) has been encoded in XML-TEI P5, including a selection of 55 documents on armament issues within Western European Union (WEU), from 1950's to 1980's.



173

UNION DE L'EUROPE OCCIDENTALE
COMMISSION INTÉRIMAIRE
GROUPE DE TRAVAIL SUR LA PRODUCTION ET LA STANDARDISATION DES ARMEMENTS

CONFIDENTIEL

PWG/E/24
Original: français
19 mars 1955

COMITE D'EXPERTS
OPPORTUNITE DE STANDARDISER

Note du Secrétaire Général

Le Secrétaire Général communique ci-joint un document soumis conjointement par les Délégations belge et française sur l'opportunité de standardiser.

Une première version de ce document a été examinée par le Comité de Rédaction à sa séance du 17 mars 1955.

Le Comité d'Experts étudiera la nouvelle version ci-annexée lors d'une de ses prochaines séances.

- 1 -

PWG/E/24

OPPORTUNITE DE STANDARDISER

1. La décision sur l'opportunité d'adopter un nouveau matériel est influencée par un certain nombre de circonstances, dont les principales sont :

(a) le programme déjà arrêté par chaque nation pour le renouvellement de son matériel et pour l'équipement de nouvelles forces,

(b) l'urgence des besoins,

(c) les disponibilités financières espérées,

(d) l'âge et la vie probable du matériel en service (dans chaque nation),

Figure 1. Transviewer. Side-by-side view digital facsimile (left) and transcription (right) (WEU sample)

In line with Booth et al. (2008), Galey and Ruecker (2010: 412-13) consider that a prototype can be the embodiment of an argument (or more), with all the key components of a "good thesis topic": to be "contestable", "defensible", and "substantive". This refers to how a prototype includes old affordances in a new way or proposes something new, to its potential of convincing people to accept it, or finally, to its intellectual and practical value.

Our idea in designing the Transviewer has been based on the following "arguments":

1. The historians or researchers in European integration studies (the CVCE's main category of readers) are always interested in comparing a transcription with the original (when available).

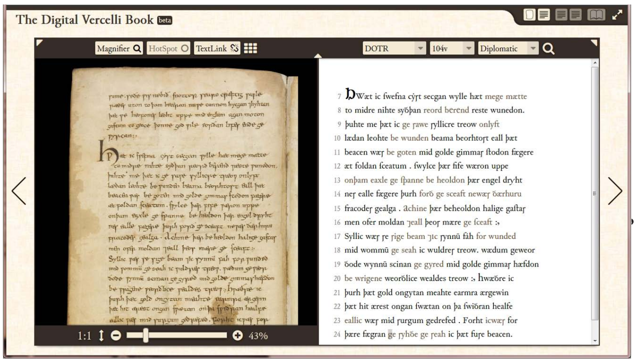
2. The architecture of the visualisation framework should be multi-project-oriented and support multiple types of historical documents (primary/secondary sources – text/image/audio/video).

A number of features to be encoded and rendered via the interface have been considered after consultations with the users (CVCE's researchers):

Transcription		
Feature type	**Encoded**	**Ignored**
Documentary	Ink colour of stamps (red, black)	Ink colour of handwritten text
Topology	Document layout (position and alignment of headers/footers/headings)	
Writing	Capitalisation and punctuation	Empty lines and exact vertical spacing on the page
Handwriting	Handwritten elements from header/headings	Handwritten fragments, sometimes not legible, from the body of the text
Textuality	Paragraphs and structural divisions	
Semantics	Named entities (e.g. names of persons/organisations/places/functions/events/products, dates)	

Table 1. "Grid of features" (Pierazzo, 2011: 467) encoded in the transcription (WEU sample)

The current version of the prototype supports functionalities such as side-by-side view of digital facsimile and transcription, page-by-page navigation, zoom-in/out, vertical scrolling, search (by names of persons/places/organisations, dates).



ƕWæt ic ſwefna cẏrt secgan wylle hæt mege mætte
to midre nihte syðþan reord berend reste wunedon.
þuhte me þæt ic ge ꞃawe ꞃyllicre treow onlyft
lædan leohte be wunden beama beorhtoꞃt eall þæt
beacen wæꞃ be goten mid golde gimmaꞃ ſtodon fægere
æt foldan ſceatum . ſwylce þær fife wæron uppe
onþam eaxle ge ſþanne be heoldon þær engel dryht
neꞃ ealle fægere þurh forð ge sceaft newæꞃ ðærhuru
fracodeꞃ gealga . ꝺchine þær beheoldon halige gaſtaꞃ
men ofer moldan ⁊eall þeoꞃ mære ge ſceaft :,
Syllic wæꞃ ꞃe ꞃige beam ⁊ic ꞃynnū fáh for wunded
mid wommū ge seah ic wuldreꞃ treow. wædum geweor
ðode wynnū scinan ge gyred mid golde gimmaꞃ hæfdon
be wrigene weorðlice wealdes treow :, ƕwæðre ic
þurh þæt gold ongytan meahte earmra ærgewin
þæt hit ærest ongan ſwætan on þa ſwiðran healfe
eallic wæꞃ mid ꞃurgum gedrefed . Forht icwæꞃ for
þære fægran ge ꞃyhðe ge ꞃeah ic þæt fuꞃe beacen.

Figure 2. EVT. Fragment (Vercelli Book)

Although the direct adaptation of EVT was considered from the beginning, different requirements for EVT and Transviewer project have been identified.

Characteristics	EVT	Transviewer
General architecture	Project-oriented (Vercelli Book manuscript)	Multi-project-oriented (documents/collections in European integration history)
XML-HTML transformation granularity	Page-oriented (one HTML file per page/manuscript image)	Document-oriented (one HTML file per XML document)

XML-HTML transformation process	HTML generation	XML transformation on the fly, in the browser, and XHTML generation
Transcription	Line-oriented	Structure-oriented (divisions, paragraphs) with semantic annotations (named entities).
Navigation	Page-by-page (supported by a single HTML file per page implementation)	Page-by-page and vertical scrolling (applied to a whole document)
Image management/loading	One by one	Images for a whole document

Table 2. EVT/Transviewer differences

Therefore, the EVT concept of side-by-side view of digital facsimile/transcription has been combined with the integration of third-party libraries (BookReader, Saxon-CE) and in-house development. The first one has been chosen for enabling on the fly loading of images, the second for supporting XSLT 2.0 transformations in the browser.

The in-house development mainly comprised modules for the implementation of a core/project specific architecture including elements of configuration, frames and buttons layout/actions, XSLT transformation and transcription rendering.

A simplified diagram of the decisions points in the first development cycle of the Transviewer is shown in Figure 4.

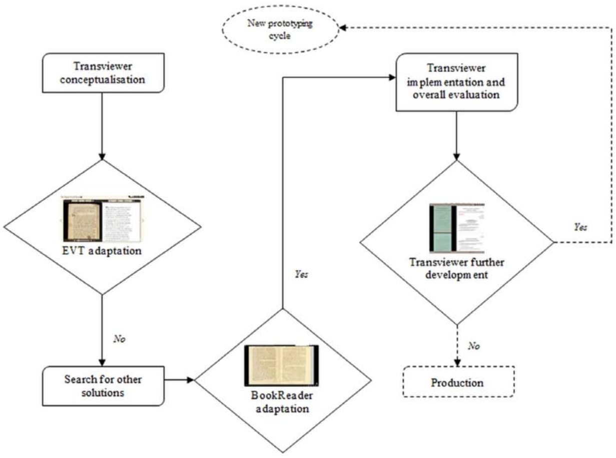
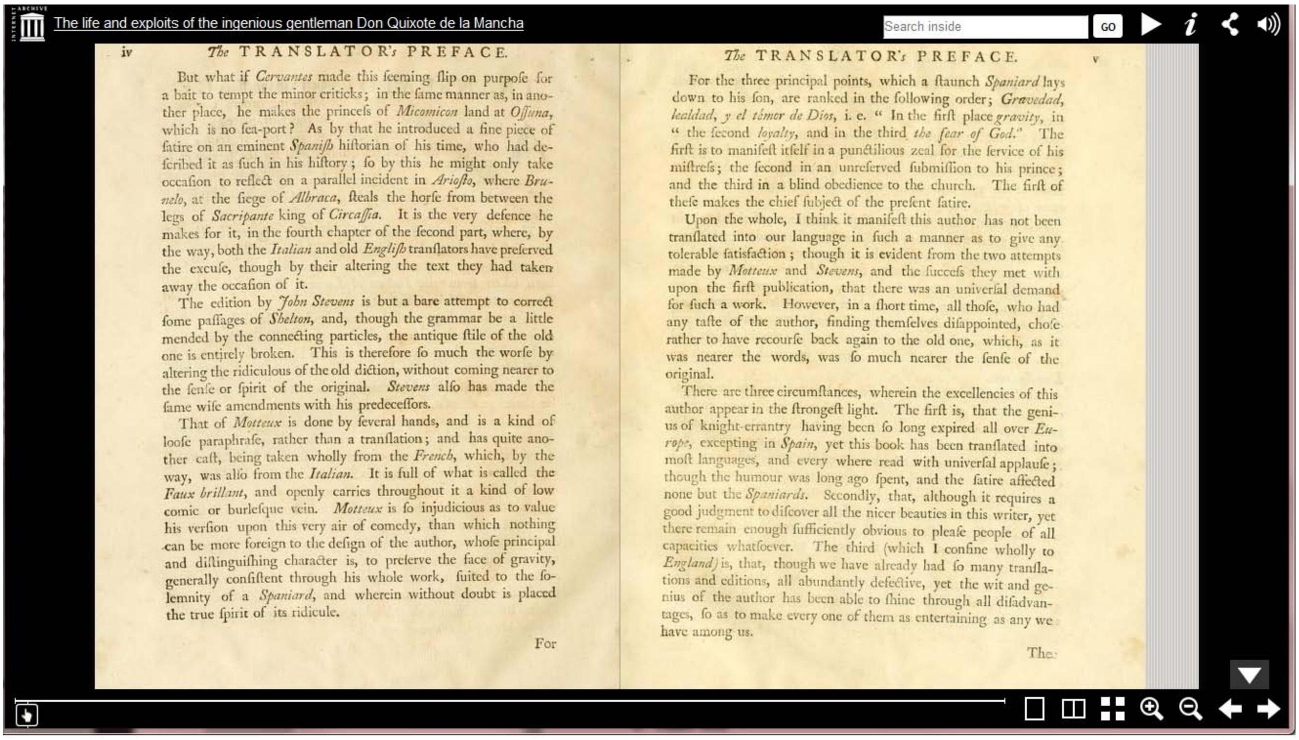

New prototyping cycle
Transviewer conceptualisation
Transviewer implementation and overall evaluation
EVT adaptation
Transviewer further development
Yes
Yes
No
No
Search for other solutions
BookReader adaptation
Production

Figure 4. Transviewer. Decision points (diamonds) in the first prototyping cycle

As illustrated, the prototyping process can imply multiple iterations. Once a functional prototype is built, the decisions on further development may reiterate similar phases of conceptualisation, search for solutions, implementation, evaluation and decision making. In this respect, our theoretical approach could be framed at the crossroad of "iterative prototyping" (Buxton and Sniderman, 1980; Buchenau and Suri, 2000; Lucena and Astua, 2012), "user-centered design" (Shneiderman and Plaisant, 2009; Warwick et al., 2009; Gibbs and Owens, 2012), and scholarly digital editions (Pierazzo, 2011; Rosselli Del Turco, 2011).


The life and exploits of the ingenious gentleman Don Quixote de la Mancha
Search inside
GO

iv *The* TRANSLATOR'*s* PREFACE.

But what if *Cervantes* made this ſeeming ſlip on purpoſe for a bait to tempt the minor criticks; in the ſame manner as, in another place, he makes the princeſs of *Micomicon* land at *Oſſuna*, which is no ſea-port? As by that he introduced a fine piece of ſatire on an eminent *Spaniſh* hiſtorian of his time, who had deſcribed it as ſuch in his hiſtory; ſo by this he might only take occaſion to reflect on a parallel incident in *Arioſto*, where *Brunelo*, at the ſiege of *Albraca*, ſteals the horſe from between the legs of *Sacripante* king of *Circaſſia*. It is the very defence he makes for it, in the fourth chapter of the ſecond part, where, by the way, both the *Italian* and old *Engliſh* tranſlators have preſerved the excuſe, though by their altering the text they had taken away the occaſion of it.

The edition by *John Stevens* is but a bare attempt to correct ſome paſſages of *Shelton*, and, though the grammar be a little mended by the connecting particles, the antique ſtile of the old one is entirely broken. This is therefore ſo much the worſe by altering the ridiculous of the old diction, without coming nearer to the ſenſe or ſpirit of the original. *Stevens* alſo has made the ſame wiſe amendments with his predeceſſors.

That of *Motteux* is done by ſeveral hands, and is a kind of looſe paraphraſe, rather than a tranſlation; and has quite another caſt, being taken wholly from the *French*, which, by the way, was alſo from the *Italian*. It is full of what is called the *Faux brillant*, and openly carries throughout it a kind of low comic or burleſque vein. *Motteux* is ſo injudicious as to value his verſion upon this very air of comedy, than which nothing can be more foreign to the deſign of the author, whoſe principal and diſtinguiſhing character is, to preſerve the face of gravity, generally conſiſtent through his whole work, ſuited to the ſolemnity of a *Spaniard*, and wherein without doubt is placed the true ſpirit of its ridicule.

For

The TRANSLATOR'*s* PREFACE. v

For the three principal points, which a ſtaunch *Spaniard* lays down to his ſon, are ranked in the following order; *Gravedad, lealdad, y el témor de Dios*, i. e. "In the firſt place *gravity*, in "the ſecond *loyalty*, and in the third *the fear of God*." The firſt is to manifeſt itſelf in a punctilious zeal for the ſervice of his miſtreſs; the ſecond in an unreſerved ſubmiſſion to his prince; and the third in a blind obedience to the church. The firſt of theſe makes the chief ſubject of the preſent ſatire.

Upon the whole, I think it manifeſt this author has not been tranſlated into our language in ſuch a manner as to give any tolerable ſatisfaction; though it is evident from the two attempts made by *Motteux* and *Stevens*, and the ſucceſs they met with upon the firſt publication, that there was an univerſal demand for ſuch a work. However, in a ſhort time, all thoſe, who had any taſte of the author, finding themſelves diſappointed, choſe rather to have recourſe back again to the old one, which, as it was nearer the words, was ſo much nearer the ſenſe of the original.

There are three circumſtances, wherein the excellencies of this author appear in the ſtrongeſt light. The firſt is, that the genius of knight-errantry having been ſo long expired all over *Europe*, excepting in *Spain*, yet this book has been tranſlated into moſt languages, and every where read with univerſal applauſe; though the humour was long ago ſpent, and the ſatire affected none but the *Spaniards*. Secondly, that, although it requires a good judgment to diſcover all the nicer beauties in this writer, yet there remain enough ſufficiently obvious to pleaſe people of all capacities whatſoever. The third (which I confine wholly to *England*) is, that, though we have already had ſo many tranſlations and editions, all abundantly defective, yet the wit and genius of the author has been able to ſhine through all diſadvantages, ſo as to make every one of them as entertaining as any we have among us.

The

Figure 3. BookReader. Fragment (Don Quixote de la Mancha)

Evaluation

Although partial evaluation had been carried out throughout the prototyping cycle, a more formal testing and evaluation phase has been conducted on the first functional version of the Transviewer. Several facets of assessment have been taken into account:

Facet	Short description	User group/ stakeholders
Technology	Technical issues have been identified and worked upon (e.g. non-uniform support for Saxon-CE in different browsers and the use in BookReader of an older version of the jQuery library). Another evaluation aspect, related to argument 2, consisted of proving the scalability of the framework by testing it with different projects samples.	CVCE's development team
User	The prototype functionalities and argument 1 have been evaluated via usability tests (Nielsen, 2000; Lund, 2001) mainly enquiring on the ease of use, ease of learning, usefulness and user satisfaction (on a scale from 1 to 5), and on suggestions for potential improvement.	Internal/external researchers as end-users of the Transviewer
Impact	Beside the impact on the final product (e.g. how many/in what proportion the functionalities/features/arguments of the prototype are reflected in the final product), another point of interest concerns its impact on other products, in particular, the initial models having inspired it.	Cooperation partners, research infrastructures

Table 3. Transviewer evaluation facets

Discussion

As an outcome of the analysis on the first prototype cycle, a new iteration and development reassessment is currently ongoing. The analysis helped us to: (1) identify and amend technical (see Table 3) and design issues (e.g. functionalities not quickly accessible or requiring extra-effort, unclear terminology or functionalities hierarchy, simplification); (2) adapt the encoding to support a larger variety of objects to be visualised (e.g. transcription only, facsimile only, transcription and audio/video); (3) better understand the needs of the users while dealing with historical documents (e.g. "trust" and "contextualisation" seem to play an important role).

Moreover, the development of new technical tools poses not only challenges to the process itself and the methods it introduces or refines but it also raises the question of sustainability: project based funding stimulates the development of new tools between different entities but far too often, this development stops once funding runs out. In the context of the informal and not project-funded cooperation between the CVCE and the EVT team, we therefore want to explore the trials and tribulations of building and maintaining a shared codebase that would be beneficial for both entities even though our specific use cases differ. As a result of the experiences made in this ongoing process of building the new version of the EVT framework, we will take the opportunity to discuss the practical challenges, opportunities and limitations of open-source development models for digital humanities projects as an approach to achieve sustainability.

In other words, we are not just testing an application but a development framework which implies considering complex theoretical-practical elements: Human-Computer Interaction (HCI) aspects, Information Technology (IT) methods and tools, specific types of documents to be published, projected audience, collaborative strategies and sustainability, etc.

Conclusion

The article proposes an analytical approach to the prototyping of a visualisation framework for historical documents. Its main assumption is that a critical perspective can be applied not only to a finished digital artefact but also to the process of its creation. In line with more traditional methods of criticism from the Humanities, a "thinking through making" viewpoint may bring into light theoretical and practical aspects related to the construction of digital tools and to the mechanisms of the "laboratory". A prospect that, by its nature, positions itself at the crossroad of the old and the new, of the past and the future.

Bibliography

BookReader. https://openlibrary.org/dev/docs/bookreader (accessed 29 February 2016).

Booth, W. C., Colomb, G. G., Williams, J. M. (2008). *The Craft of Research*. Chicago & London: The University of Chicago Press.

Buchenau, M., Suri, J. F. (2000). Experience Prototyping. New York: *DIS '2000*. https://www.ideo.com/images/uploads/news/pdfs/FultonSuriBuchenau-Experience_PrototypingACM_8-00.pdf (accessed 29 February 2016).

Buxton, W., Sniderman, R. (1980). Iteration in the Design of the Human-Computer Interface. *Proceedings of the 13th Annual Meeting, Human Factors Association of Canada*, pp.

72-81. http://echo.iat.sfu.ca/library/buxton_80_Iteration.pdf (accessed 29 February 2016).

CVCE (Centre Virtuel de la Connaissance sur l'Europe). http://www.cvce.eu/ (accessed 29 February 2016).

EVT (Edition Visualization Technology). http://sourceforge.net/projects/evt-project/ (accessed 29 February 2016).

Galey, A., Ruecker, S. (2010). How a prototype argues. *Literary and Linguistic Computing*, **(25)**4: 405-24.

Gibbs, F., Owens, T. (2012). Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs, *Digital Humanities Quarterly*, **(6)**2. http://digitalhumanities.org/dhq/vol/6/2/000136/000136.html (accessed 29 February 2016).

Lucena, B., Astua, F. (2012). Iterative prototyping and rapid service design user evaluation. *Participatory Innovation Conference 2012*, Melbourne, Australia. www.pin-c2012.org/, http://www.academia.edu/4074261/Iterative_prototyping_and_rapid_service_design_user_evaluation (accessed 29 February 2016).

Lund, A. M. (2001). Measuring Usability with the USE Questionnaire. *STC Usability SIG Newsletter*, **(8)**:2. http://garyperlman.com/quest/quest.cgi?form=USE(accessed 29 February 2016).

Manovich, L. (2007). Q and A Session at the *Digital Humanities (DH) 2007 Conference*, Urbana-Champaign, IL, (cited in Galey and Ruecker, 2010).

Nielsen, J. (2000). Why You Only Need to Test with 5 Users, *Nielsen Norman Group*, March 19, 2000. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ (accessed 29 February 2016).

Pierazzo, E. (2011). A rationale of digital documentary editions. *Literary and Linguistic Computing*, **(26)**4: 463-77.

Rosselli Del Turco, R. (2011). After the editing is done: Designing a Graphic User Interface for digital editions. *Digital Medievalist*, 7. ISSN: 1715-0736. http://www.digitalmedievalist.org/journal/7/rosselliDelTurco/ (accessed 29 February 2016).

Rosselli Del Turco, R., Buomprisco, G., Di Pietro, C., Kenny, J., Masotti, R., Pugliese, J. (2014-2015). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. In *Journal of the Text Encoding Initiative*, Issue 8 - PREVIEW | 2014-2015: Selected Papers from the 2013 TEI Conference ,https://jtei.revues.org/1077v(accessed 29 February 2016).

Saxon-CE, http://www.saxonica.com/ce/user-doc/1.1/index.html (accessed 29 February 2016).

Shneiderman, B., Plaisant, C. (2009). *Designing The User Interface: Strategies for Effective Human-Computer Interaction*, Addison Wesley Pub Co Inc.

- **TEI** (Text Encoding Initiative). http://www.tei-c.org/index.xml (accessed 29 February 2016).

Warwick, C., Fisher, C., Terras, M., Baker, M., Clarke, A., Fulford, M., Grove, M., O'Riordan, E., Rains, M. (2009). iTrench: A study of user reactions to the use of information technology in field archaeology, *Literary and Linguistic Computing*, **(24)**2: 211-23.

Combining Corpora and Statistics using Geographical Technologies: New Evidence on Nineteenth Century Infant Mortality Decline in England and Wales

Paul Atkinson
p.atkinson3@lancaster.ac.uk
Lancaster University, United Kingdom

Ian Gregory
I.Gregory@lancaster.ac.uk
Lancaster University, United Kingdom

Catherine Porter
c.porter2@lancaster.ac.uk
Lancaster University, United Kingdom

The division between quantitative and qualitative approaches is fundamental to the study of the past. Quantitative approaches are predominantly used to study statistical sources in fields such as historical demography and economic history where large numerical databases are available. These approaches are well suited to situations where large volumes of digital statistics are available and are very good at finding relationships between variables. They can be criticised on two levels: firstly, while good at identifying relationships they are poor at establishing the causal mechanisms that cause them. Secondly, and more fundamentally, most information about the past is not in numerical form and thus cannot form part of a traditional quantitative analysis, thus many relevant factors cannot be included within the analysis. Qualitative sources, particularly texts, are richer in both the range of material available and in the amount of detail it provides about the conditions in which people lived. They are however much more complex to work with and traditionally required close reading which is slow and selective. Digital technologies offer the potential to overcome this divide and make use of the combined advantages of both types of source, making use of statistical sources to identify patterns and relationships, and textual sources to help explain the patterns found. This paper presents an example of this based on infant mortality in Victorian and Edwardian England and Wales.

Infant mortality refers to deaths before the age of one, usually expressed as an infant mortality rate (IMR) of infant deaths per thousand births. An orthodox story is that infant mortality decline was brought about by government action on sanitation (Szreter, 1991; Woods et al., 1988; Woods and Shelton, 1997). More recently Gregory (2008) used a GIS database and demonstrated that infant mortality decline actually started before the Public Health Acts, that decline was earlier and steeper in rural areas of the south and east of England than anywhere else, and that rural parts of the north, Wales and the West Country had among the lowest rates of improvement. He was, however, unable to explain this variation with the data available. This paper uses a combination of statistical and textual sources to attempt to provide explanations for infant mortality decline with an emphasis on rural areas.

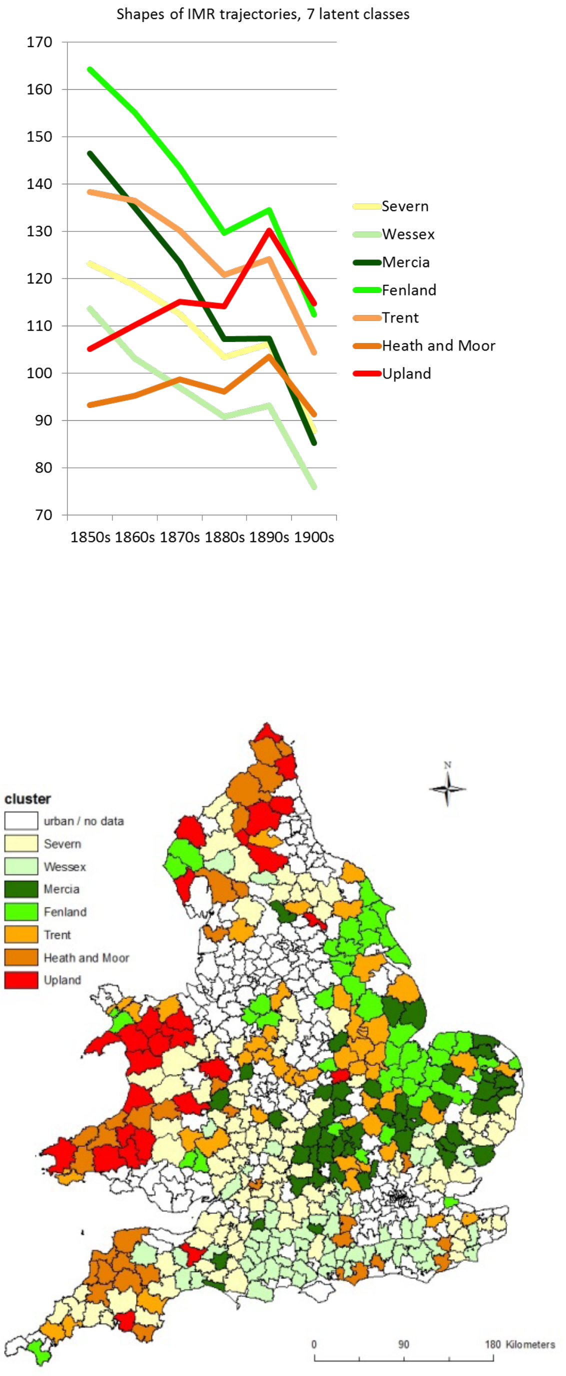


Figure 1: Temporal trajectories and geographical locations of the seven latent classes

In the quantitative stage, as many geographically disaggregate independent variables as possible were assembled to explore what other factors seemed to be associated with infant mortality decline. Analysing change over time and space for multiple variables is difficult. A technique called Latent Trajectory Analysis to do this but combine it with the use of GIS (Nagin, 1999). Figure 1 shows how this techniques was used to group districts into seven clusters based on their temporal characteristics, and then map where rural districts with similar trajectories are found. This confirms that there was a clear geography to infant mortality decline. The biggest improvements occurred in rural areas in the south and east in the clusters described as 'Fenland' and 'Mercia.' Areas showing the lowest declines, which in some cases actually got worse, were in the north and west in the clusters described as 'Upland' and 'Heath and Moor.'

Further analysis reveals statistically significant relationships between infant mortality change and rates of female TB and a less strong relationship with female literacy. These relationships have been shown elsewhere in the literature, however our findings show that these factors were important in rural areas as well as in urban ones. Interestingly convincing relationships could not be found with population density or fertility, these would be expected from the literature. The other striking finding was that the most important independent variable in every model was 'decade', in other words, change over time that none of our other variables could account for. Fundamentally this tells us three things: first that there are some very interesting geographical patterns to infant mortality decline, second that this is related to some other factors, particularly female health and education, and third that even though we have assembled what we believe to be as many quantitative independent variables as it is practical to use, these are unable to account for much of the change in infant mortality over this period.

The relative lack of explanatory ability by the quantitative analysis is hardly surprising given the lack of variables on a whole range of topics that may be associated with improving infant health, including: sanitary conditions, access to midwives and healthcare, attitudes towards breastfeeding, access to safe cows' milk, and so on which have never been captured in statistical form. Thus, if we are going to improve our understanding we need to explore non-quantitative sources. Of particular interest here is the British Library's Nineteenth Century Newspapers collection, a corpus of over 50 newspapers, most of which are available for series that cover most of the century. The corpus is at least 30 billion words although, to date we have been working with individual newspaper series which are usually hundreds of millions of words.

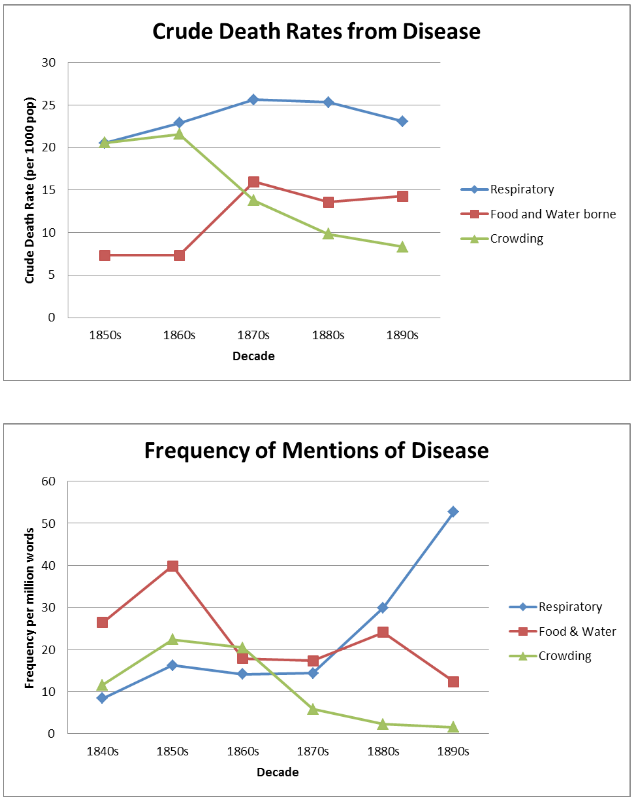


Figure 2: Crude death rates from disease classes and frequency of instances of these diseases in the Era

Figure 2 shows the use of one newspaper, the Era, to explore the relationship between interest in a variety of diseases associated with the young, classed according to whether they are respiratory diseases, food and water-bourn diseases, and diseases associated with crowding (Woods, 2000). The frequency with which these diseases are mentioned in the corpus is compared with the death rates from the diseases. Different disease types show quite different results. Interest in diseases of crowding follows the decline in these diseases, particularly associated with the decline of scarlet fever and typhus. Food and water death rates rise overall while interest in them seems to fall, while deaths from respiratory diseases are broadly flat but there is a major increase in interest in them in the 1880s and 1890s. We can then explore the collocates to these terms, the words that are found close to them. This shows four major classes of collocates: first, other words associated with disease, included other disease names, and words such as 'died', 'attack' and 'epidemic.' Second there are symptoms of disease, third are brand names of medicines typically found in advertisements, and fourth there are military terms which are found in the 1850s and are related to the Crimean War.

To allow us to explore the geographies within themes within the newspapers we have developed a technique we call concordance geoparsing (Rupp et. al., 2014). This involves firstly extracting concordance lines around the search-term. These concordance lines are then geo-parsed using the Edinburgh Geoparser (Grover et. al., 2010) which identifies place-names and matches them to a gazetteer to allocate them to a grid reference. The results are then explored to check for errors which are corrected with the corrections added to an updates file. In this way we can cumulatively remove geoparsing errors and have confidence in its results.

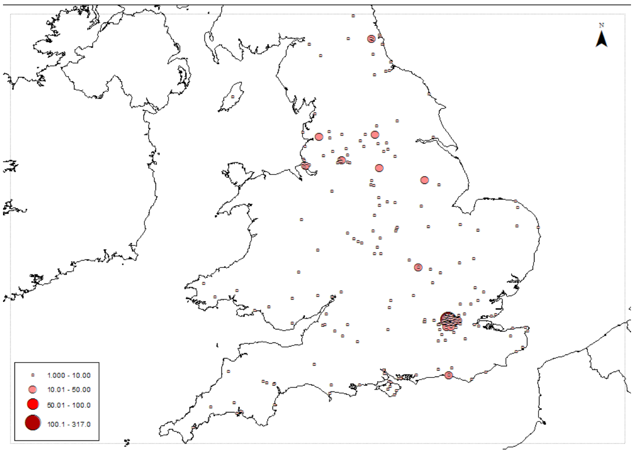


Figure 3: Locations associated with disease instances from the Era

Figure 3 shows the place-names associated with the diseases in the Era. It is noticeable that these are concentrated in urban areas but there are exceptions to this. We will present comparisons between the number of references to a disease and the number of deaths from it in a district, showing how media interest was not related to disease severity.

Taking this work further, we will take regional newspapers from rural areas that experienced the highest and lowest declines and explore these for collocates and places that are associated with diseases themselves and a wide range of potential causal factors that may be related to them. In this way we will be combining the descriptive strengths of statistical data with the explanatory power of textual sources, using text to work around insufficiencies of data and making use of the temporal and geographical detail in both.

Acknowledgement

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850).

Bibliography

Gregory, I. N. (2008). Different places, different stories: Infant mortality decline in England and Wales. *1851-1911 Annals of the Association of American Geographers* **98**: 773-94.

Grover, C., Tobin, R., Woollard, M., et al. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A* 368, pp. 3875-889.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, **4**: 139-57.

Rupp C. J., Rayson P., Gregory I., et al. (2014). Dealing with heterogeneous big data when geoparsing historical corpora. *Proceedings of the 2014 IEEE Conference on Big Data*, pp. 80-83.

Szreter, S. (1991). The GRO and the Public-Health Movement in Britain, 1837-1914. *Social History of Medicine*, **4**: 435-63.

Woods, R. I. (2000). *The demography of Victorian England and Wales*. Cambridge: Cambridge University Press.

Woods, R. I., Watterson, P. A. and Woodward, J. H. (1988). The causes of rapid infant mortality decline in England and Wales, 1861-1921. Part I. *Population Studies*, **42**: 343-66.

Woods, R. I. and Shelton, N. (1997). *Atlas of Victorian Mortality*. Liverpool: Liverpool University Press.

vWise: Visual Workspace for Information Seeking and Exploration

Michael Neal Audenaert
neal@idch.org
Texas Center for Applied Technology, United States of America

Matthew Barry
mbarry@tamu.edu
Texas Center for Applied Technology, United States of America

Paul Bilnoski
bilnoski@tamu.edu
Texas Center for Applied Technology, United States of America

Introduction

Increasingly, scholars have access to large, heterogeneous collections of information such as those provided by the Digital Public Library of America, Europeana, or Canadiana as well as more targeted thematic research collections like the Whitman Archive. In additions to purpose-built humanities projects, scholars commonly make use of general-purpose resources such as Flickr, Google Search, HathiTrust Digital Library and many other services.

Traditional search interfaces allow scholars to rapidly search for specific items and to explore a collection via facetted search interfaces and other techniques. Search, however, is only one part of a complex ecosystem of behaviors associated with information seeking (Toms and O'Brien, 2008; Buchanan et al., 2005). In practice, scholarly information seeking is not characterized by a single search event or interaction with a single system. Rather, it is a process that takes place over an extended period of time and involves searching different sources for potentially relevant material. As material is found, scholars organize, annotate and make notes about the retrieved material; activities that lead to more questions and more seeking.

These tasks serve a dual purpose of cataloging material for re-discovery, use and interpretation as well as engaging the participant in a process of sensemaking and internalization. The fundamental purpose of this work is not merely the discovery, consumption or indexing of information or even the production of a concrete research output such as a paper. Instead, scholarly information seeking serves to hel

p the seeker develop an internalized, systematic understanding of a body of knowledge and cultivate a distinctive interpretive voice (Audenaert and Furuta, 2010).

We are developing the Visual Workspace for Information Seeking and Exploration (vWise[1]) to support these tasks. This paper introduces the key features and capabilities of the vWise platform and discusses how those features and capabilities reflect and build on a theoretical framework for designing systems to support open-ended information seeking and exploration.

System Overview

vWise is an extensible Web-based application framework that allows scholars to search for content from different data providers and bring the resulting information together into a single display. Once resources have been added to the workspace, they can then organize and reframe that information, for example by culling unwanted results or juxtaposing digital surrogates for newly discovered relationships.

Panels are the primary way people interact with content in the vWise interface. They are used to display different types of information drawn from different sources in a single workspace. Users can organize this information by rearranging panels within the workspace and add visual annotations by setting properties such as background color, border style, font properties and drop shadows. The panels themselves support user interactions tailored to the types of content they display. For instance, users could read a book from HathiTrust, watch a video from YouTube, explore metadata from DPLA or manipulate manuscript page images to improve readability.

Users have several options for adding content panels to a workspace. A basic search service allows them to execute searches using pluggable content providers. Alternatively, custom panel implementations may provide advanced searching capabilities or allow users to drag composite content such as pages from a manuscript onto the workspace to create new panels.

Workspaces provide a site for both individual and collaborative work. To enable collaboration, a user can share a workspace with others. Modifications to this shared workspace are reflected on all users' display simultaneously. This allows both synchronous and asynchronous collaboration.

As a framework, vWise is designed to be customized and configured by an application developer prior to being deployed. This customization allows the basic application to be extended in four major ways:

Integration with external data sources. vWise provides a pluggable architecture for defining interfaces to external data sources such as a digital library or search service. For instance, we provide data sources that allow users to load videos from YouTube videos, search Wikipedia and display web-pages.

Custom panels for working with data objects. Individual units of content are displayed using panels that provide basic support for moving, resizing and styling. Extensions support rich interactions tailored to specific types of content. For example, an image annotation tool could allow the creation and storage of annotations.

Ad hoc panels for interacting with underlying search services. The core use-case for panels is to display and organize individual content elements. Panels can also be used to implement other application services. For example, we envision implementing a search service panel that will display basic and advanced search options, display facets from the currently active search, maintain a history of recent search and display search results as panels in the main workspace.

Integration with server-side workspace persistence. vWise runs in a browser and stores workspaces and workspace configuration information via a well-defined RESTful interface. Application developers can use the default implementation that comes with vWise or provide a custom implementation. For instance, we developed a custom data storage implementation to connect the vWise interface with a proprietary system used to train emergency responders.

Supporting Information Seeking and Exploration

Information seeking and exploration in scholarly research is an intensive, creative activity. Supporting these activities require tools that go beyond merely helping scholars find resources to provide environments that reflect and facilitate the creative process. vWise is based on ideas that have emerged from the literature in hypermedia systems, digital libraries and creativity research, including the following core concepts:

Information Triage: The process of information seeking requires people to rapidly assess the utility of various information resources, discard those that aren't relevant and prioritize those that show promise (Marshall and Shipman, 1997). vWise goes beyond traditional search interfaces by allowing users to remove individual results, prioritize the remaining objects and combine the results from multiple searches in a single workspace.

Incremental Formalization: One aspect of open-ended information seeking tasks is sense-making—the process of gaining a broad understanding of the structure of ideas within a domain. During sense-making, the emerging organizational framework is partial, provisional and implicit (Shipman and Marshall, 1999). vWise facilitates incremental formalization of knowledge by allowing users to express an emerging organization framework implicitly via the spatial arrangement and visual properties of the panels displayed on the workspace. Within this environment, user can rapidly form and reform organizational structures by manipulating the visual characteristics of the display.

Representational Talkback: The visually expressed knowledge structure serves as an externalized representation of a user's internal, evolving mental models. Creating and interacting with this content supports representational talkback (Schön, 1983; Nakakoji et al., 2000). Representational talkback occurs when material externalizations of an internal mental model initiate a "reflective conversation" in which they talk back to their creator to inform subsequent stages of the sense making process.

Heterogeneous Sources and Material: Complex information seeking tasks rarely involve one-stop shopping. vWise allows people to integrate, analyze and synthesize information from different sources, by bringing content into a unified information organization space. Manipulation of spatial and visual properties (border color, drop-shadows, backgrounds, etc.) provides a lightweight and open-ended interface manipulating this content. Domain-specific panel implementations allow users to engage in content-specific interactions.

Summary

vWise provides support for the complex information seeking needs of scholars and other professionals that goes beyond the capability of traditional search systems. It allows people to gather information from multiple sources and to work either independently or collaboratively to organize and analyze that material using a workspace metaphor.

To date, we have demonstrated prototype implementations of vWise in two different settings. While initially conceived and designed to support the needs of scholars in the humanities, we have integrated a version of vWise into the Emergency Management*Exercise System (EM*ES). EM*ES is a simulation tool used by Texas A&M Engineering Extension Service (TEEX) to train incident managers, supervisors, and jurisdiction officials in the management of a large-scale crisis (TEEX, 2016). We have added vWise as a component to support the training of communication and coordination between emergency managers and cyber response teams. vWise is currently in use in a series of training exercises starting in January of 2016 and running through September 2017. During this period we will investigate options for more widespread use within the EM*ES application.

Our second deployment of vWise was a demonstration prototype for use in wall-sized display of the Humanities Visualization Space (HVS) at Texas A&M University's Initiative for Digital Humanities, Media and Culture (IDHMC). While vWise is envisioned primarily for desktop-oriented use, when working on knowledge organization tasks, bigger is often better. The HVS provides just such a space and opens intriguing possibilities for scholars to work with this class of interfaces both as individuals and in collaborative settings.

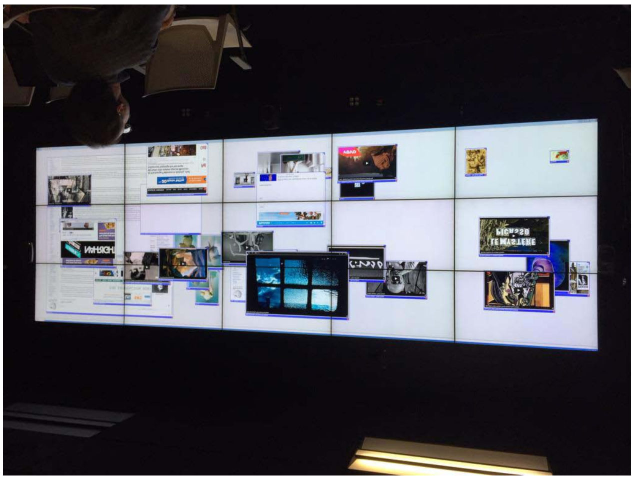


Figure 1: vWise demonstration at the IDHMC Humanities Visualization Space

Moving forward, we continue to develop the core implementation of vWise framework and to add additional content provider connections and panel implementations for displays. We are specifically interested in developing integrations with major content providers for humanities and cultural heritage scholars and are working to pursue such collaborations.

Bibliography

Audenaert, N. and Furuta, R. (2010). What Humanists Want: How Scholars Use Source Materials. *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. (JCDL '10). New York, NY, USA: ACM, pp. 283–92. doi:10.1145/1816123.1816166 (accessed 13 March 2016).

Audenaert, N., Lucchese, G. and Furuta, R. (2010). CritSpace: A Workspace for Critical Engagement within Cultural Heritage Digital Libraries. In Lalmas, M., Jose, J., Rauber, A., Sebastiani, F. and Frommholz, I. (eds), *Research and Advanced Technology for Digital Libraries*. (Lecture Notes in Computer Science 6273). Springer Berlin Heidelberg, pp. 307–14. http://link.springer.com/chapter/10.1007/978-3-642-15464-5_31 (accessed 13 March 2016).

Buchanan, G., Cunningham, S. J., Blandford, A., et al. (2005). Information Seeking by Humanities Scholars. In Rauber, A., Christodoulakis, S. and Tjoa, A. M. (eds), *Research and Advanced Technology for Digital Libraries*. (Lecture Notes in Computer Science 3652). Springer Berlin Heidelberg, pp. 218–29. http://link.springer.com/chapter/10.1007/11551362_20 (accessed 13 March 2016).

Toms, E. G. and O'Brien, H. L. (2008). Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation*, **64**(1): 102–30. doi:10.1108/00220410810844178.

Marshall, C. C. and Shipman, F. M., III (1997). Spatial Hypertext and the Practice of Information Triage. *Proceedings of the Eighth ACM Conference on Hypertext*. (HYPERTEXT '97). New York, NY, USA: ACM, pp. 124–33. doi:10.1145/267437.267451 (accessed 13 March 2016).

Nakakoji, K., Yamamoto, Y., Takada, S., et al. (2000). Two-dimensional Spatial Positioning As a Means for Reflection in Design. *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*. (DIS '00). New York, NY, USA: ACM, pp. 145–54. doi:10.1145/347642.347697 (accessed 13 March 2016).

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.

Shipman, F. M. and Marshall, C. C. (1999). Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work (CSCW)*, **8**(4): 333–52. doi:10.1023/A:1008716330212.

TEEX (2016). TEEX | Emergency Operations Training Center (EOTC). https://teex.org/Pages/about-us/emergency-operations-training-center.aspx (accessed 13 March 2016).

Notes

[1] vWise builds on key concepts and findings of a previous project, CritSpace (Audenaert et al., 2010) and provides a new implementation and enhanced capabilities.

From Index Cards to a Digital Information System: Teaching Data Modeling to Master's Students in History

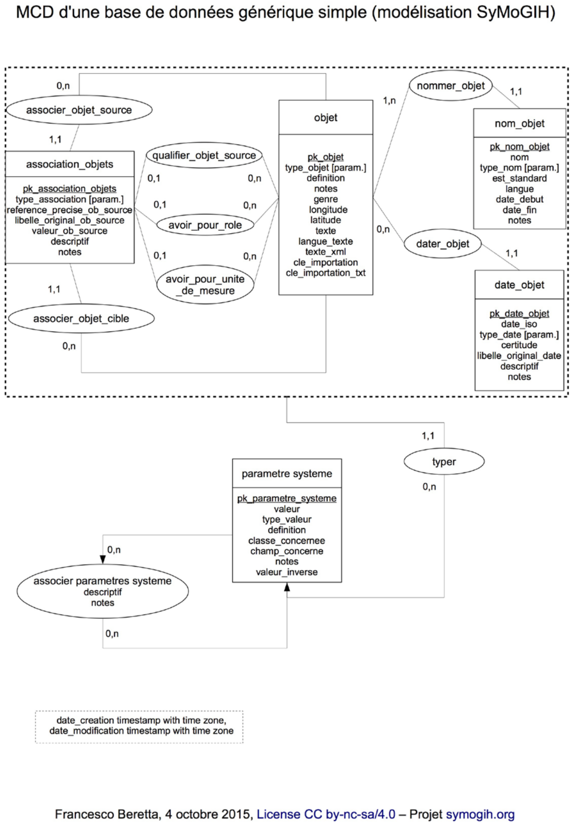
Francesco Beretta
francesco.beretta@ish-lyon.cnrs.fr
CNRS - Université de Lyon, France

For the last five years I've taught a course on "computer science for historians" at University Lyon3 Jean Moulin. The course includes 20 hours in ten sessions and has attracted twenty to forty students each year. It addresses history students during the first semester of their master studies: they start at this stage with information collection in archival sources and bibliography, which they can later exploit to write their master's thesis. Thus, the aim is to provide them with methods and digital tools for modeling and storing information, and then for subjecting it to interrogation, visualization and analysis. This is a great challenge because many students still use paper for taking notes when they analyze historical sources and are not used to working with software that is not completely self-installing. Furthermore, students will receive from their tutors all kinds of research subjects, from Ancient to Modern history, and they often want to analyze quite complex information which one cannot store in a simple spreadsheet. For this reason, the pedagogical challenge is also a challenge for the digital humanist: how can the students be provided with a generic and flexible information system of ready use for their research but sophisticated enough to store any sort of data?

In this paper I will treat some theoretical and practical aspects of the digital information system I devised to cope with this problem, and will present some issues raised by recourse to the information system that concern both students and teacher. The manuscript of the course is publicly available on a dedicated website[1]: anyone interested can download the tools I developed and test the methods proposed to the students, or employ them for their own teaching. As I teach in French the documentation, interfaces, etc. are written in this language. The information system I propose in the course combines the experience acquired in developing the symogih.org project[2], a collaborative platform for storing and sharing structured historical data, with the method of semantic annotation of texts adopted in our platform for digital editions[3] in accordance to the *Text encoding initiative*'s

guidelines (TEI). These data production practices, both as structured data and encoded texts, must be radically simplified to cope with the pedagogical need exposed above and this requires working on a high level of abstraction.

The first component of the information system provided to the students is a relational database designed using a generic data model[4]. In the center of the model (Figure 1), the object class, having the same sense as the “Endurant” class in DOLCE[5], or the “Persistent Item (E77)” class in CIDOC-CRM[6], comprises individual actors, institutions, places, concepts, etc. about which students will be collecting information. The function of this class is to provide an identifier for each individual, in turn characterized by one or more names, a time span of existence, a type and an accurate textual definition. The database also allows treatment of some basic associations between objects defined in a class “system parameter” – a typical component of a generic data model – whose instances are predefined by the teacher. This simplifies the use of the database by students and guides them in their first steps of data production, but if needed parameters can be extended to other kinds of relationships. A simple PHP interface is added to facilitate data capture.


MCD d'une base de données générique simple (modélisation SyMoGIH)

Francesco Beretta, 4 octobre 2015, License CC by-nc-sa/4.0 – Projet symogih.org

Figure 1

The database is implemented using PostgreSQL because this open-source database provides extended features in datatype treatment (namely XML) and comes along with a procedural language (PL/pgSQL) allowing data treatment in a SQL context without having to learn a different programming language[7]. The teacher can thus write predefined functions to help the student prepare, transform and code the data before further treatment. A spatial extension is also available (PostGIS) which permits working with geo-referenced data if needed[8]. PostgreSQL is therefore a kind of "Swiss Army knife" for historical data storage and treatment.

If the "Objects" ("Endurants" or "Persistent items") are identified in the database, where then is collected information about them? According to the *symogih.org* semantic data model[9], a "Knowledge Unit" is an atomized portion of information that expresses a relationship among objects situated in space and time, established on critical analysis of documents. The class "Knowledge Unit" is therefore equivalent to the "Temporal entity (E2)" class in CIDOC-CRM or "Perdurant" class in DOLCE: “An endurant lives in time by *participating* in some perdurant(s). For exemple, a person, which is an endurant, may participate in a discussion, which is a perdurant”[10].

In former years, "knowledge units" were also stored in the database, as the "objects" are presently[11]. Pedagogical experience has shown that the degree of abstraction required for modeling information in form of structured data is generally too steep for training digital historians, although some students used this method with ease. The newly proposed information system comprises therefore a second component which consists in a text encoding method using some specific TEI tags and attributes. These allow semantic text encoding: "knowledge units" can be directly annotated into the text, thus marking up named entities with the database identifiers of the related objects and then encoding their properties and relationships in the text with specific tags and attributes[12].

But this method raises the question of the XML editor to adopt for text semantic encoding, meaning the addition of a further software component to the workflow of data production providing XML schema validation and also tools for querying the encoded text. XML text encoding is more suitable and I prefer it for PhD student and researcher training, but this demands a supplementary specific instruction that it impossible to provide in the limited master's course time. I therefore conceived a way of semantically tagging the text in a simple text editor or word processing program using curly brackets instead of angle brackets and replacing XML-attributes by predefined codes. This method is described on the course wiki that also furnishes instructions for using regular expressions for proper encoding[13]. Regular expressions are then used in a PL/pgSQL script in the database to transform the curly brackets and their content into real XML tags and attributes: the encoded tag “{en2ai_10}Johannes Kepler{/en}” becomes “<en type=“ai” ref=“2” ana=“10”>Johannes Kepler</en>” (belonging to a course-specific namespace). This transformation allows storage of the encoded text in a PostgreSQL XML field and consequently benefit of the full power of the XPath and SQL queries, and programming capabilities of PL/pgSQL, to extract information from the texts.

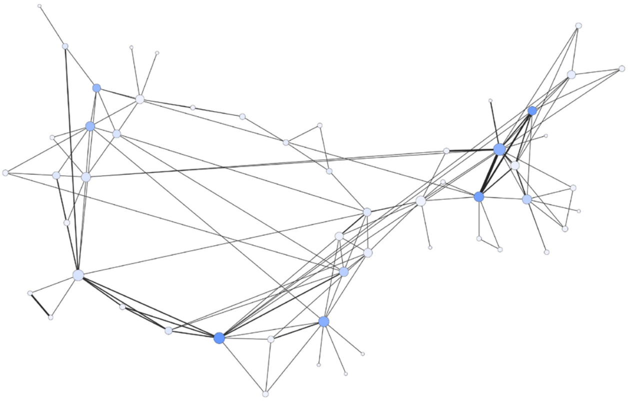


Figure 2

The workflow of data production and treatment ends with the phase of data analysis and visualization. For this purpose I adopted the R software that can be directly connected to a PostgreSQL database and provides many useful libraries. For instance, a former student produced data about relationship between persons attested by medieval charters that can be used for network analysis (Figure 2).

The students can send the teacher a dump of their database and formulate the research questions that the latter will transcribe into SQL, XPath or procedure language queries for extracting data, before sending this back to the students. Building upon these examples the students can themselves adapt the queries and scripts to new research questions. A wiki dedicated to each student's project can be created to document the specific workflow of each research project: it is not public but it is accessible to all other students participating in the process of data production and analysis. The students can use the results of data analysis and visualization to formulate new research hypotheses or they can integrate them into their master's thesis.

In this paper I will present the essential conceptual and technical aspects of the whole workflow and consider three major advantages of this pedagogical approach for the disciplinary domain of digital history. First, students gain the experience of managing a workflow going from installation and personal practice on a solid community maintained open-source software, to reflection on data modeling concerning their own research agenda, to collaborative data and project management through a wiki, to an introduction in data mining and visualization techniques. Secondly, the abstraction level of the data model and text encoding practice proposed to the students implicitly introduces them to knowledge management and data production according to present-day standards like CIDOC-CRM and *Text encoding initiative*: from this perspective historical knowledge is modeled as a graph of objects situated in time and space and linked to the texts from which they derive. Thus —and this is the third advantage— the course acquaints students with the basic principles of linked data and of semantic text encoding, introducing them to the concepts and practice of resource sharing and data curation: the datasets I use for the exercises come from the French national library (BNF) SPARQL endpoint and DBPedia, and the texts from Wikipedia. In a final part, I will discuss the issues that this pedagogical approach raises for master's students in history.

Bibliography

Alerini, J. and Lamassé, S. (2011). Données et statistiques. L'avenir du travail en ligne de l'historien. In: Genet, J.-P. and Zorzi, A. (eds), *Les historiens et l'informatique. Un métier à réinventer.* Rome: Ecole française de Rome, pp. 171-187.

Beretta, F. (2015). The symogih.org project and TEI : encoding structured historical data in XML texts. In: *Text Encoding Initiative Conference and Members' Meeting 2015. Connect,* *Animate, Innovate*, Lyon, France: https://halshs.archives-ouvertes.fr/halshs-01251915v1 .

Cellier, J. and Cocaud, M. (2001). *Traiter des données historiques: méthodes statistiques, techniques informatiques*. Rennes: Presses universitaires de Rennes.

Eide, Ø. (2014-2015). Ontologies, Data Modeling, and TEI. *Journal of the Text Encoding Initiative,* **8:** Selected Papers from the 2013 TEI Conference. http://jtei.revues.org/1191.

Erickson, A. T. (2013). Historical Research and the Problem of Categories. In: Dougherty, J. and Nawrotzki, K. (eds), *Writing History in the Digital Age.* Ann Arbor: University of Michigan Press, pp. 133-145.

Gast, H., Leugers, A. and Leugers-Scherzberg, A. H. (2010). *Optimierung historischer Forschung durch Datenbanken. Die exemplarische Datenbank "Missionsschulen 1887-1940"*. Bad Heilbrunn: Verlag Julius Klinkhardt.

Jordanous, A., Stanley, A. and Tupman, C. (2012). Contemporary transformation of ancient documents for recording and retrieving maximum information: when one form of markup is not enough. Proceedings of Balisage, **8:** The Markup Conference 2012, http://www.balisage.net/Proceedings/vol8/html/Jordanous01/BalisageVol8-Jordanous01.html.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2003). *WonderWeb Deliverable D18 Ontology Library (final).* Trento: Laboratory For Applied Ontology. PDF version: http://wonderweb.man.ac.uk/deliverables.shtml .

Notes

[1] http://phn-wiki.ish-lyon.cnrs.fr/doku.php?id=td_histoire_numerique:accueil(all websites were accessed on 30 October 2015).

[2] Documentation about the project and links to its the different components are provided on the project main website : http://symogih.org/ .

[3] http://xml-portal.symogih.org/web-publications.html .

[4] Cf. https://en.wikipedia.org/wiki/Generic_data_model. Databases are used in historical research both at individual and project level. See, e.g., Gast, Leugers and Leugers-Scherzberg, 2010; Alerini and Lamassé, 2011; Cellier and Cocaud, 2001; Erickson, 2013. The novelty of the method proposed by the symogih.org project, and in my course, is the abstraction level allowing to treat any kind of historical information.

[5] DOLCE : a Descriptive Ontology for Linguistic and Cognitive Engineering: http://www.loa.istc.cnr.it/old/DOLCE.html.

[6] CIDOC Conceptual Reference Model (CRM): http://www.cidoc-crm.org/.

[7] http://www.postgresql.org/docs/9.4/static/plpgsql.html.

[8] http://postgis.net/.

[9] http://symogih.org/?q=rdf-publication .

[10] Cf. Masolo et al., 2003: 14.

[11] http://phn-wiki.ish-lyon.cnrs.fr/doku.php?id=td_histoire_numerique:installation_db_2013 .

[12] Some pages of the symogih.org project's user manual provide the encoding specification for XML/TEI semantic annotated texts using the symogih.org ontology: https://groupes.renater.fr/wiki/symogih/symogih_manuel/edition_de_textes_en_xml-tei. This method was presented at the TEI 2015 conference in Lyon, cf. Beretta 2015. The Special Interest Group Ontologies in the TEI Consortium is devoted to this approach. See the GIS Ontologies wiki : http://wiki.tei-c.org/index.php/SIG:Ontologies and Eide, 2014-2015. A similar approach is represented by Jordanous, Stanley and Tupman, 2012.

[13] http://phn-wiki.ish-lyon.cnrs.fr/doku.php?id=td_histoire_numerique:exercice_2.

Modelling Taxonomies of Text Reuse in the Deipnosophists of Athenaeus of Naucratis: Declarative Digital Scholarship

Monica Berti
monica.berti@uni-leipzig.de
University of Leipzig, Germany

Mary Daniels
ellie.daniel@furman.edu
Furman University, USA

Samantha Strickland
sami.strickland2@furman.edu
Furman University, USA

Kimbell Vincent-Dobbins
kimbell.vincent-dobbins@furman.edu
Furman University, USA

This paper presents work on documenting text reuse of fragmentary authors and of extant works. By **fragmentary** we mean authors whose texts are lost and known through quotations and references by other authors. Within ancient Greek literature 60% of authors is preserved only in fragments, showing the challenge of working with innumerable pieces of reuse scattered in our textual heritage (Berti et al., 2009). This work is necessarily prior to any specific research questions. We cannot inquire into, e.g., the historical works of Istrus the Callimachean until we can comprehensively and precisely catalogue the surviving fragments of Istrus; nor can we ask "how did intellectuals in the 3rd century CE read epic poetry?", until we can comprehensively identify instances of Homeric text reuse and work with them in their context.

The term **fragment** is the result of print editorial practices, where chunks of text preserving traces of lost authors and works are extracted from their contexts and reprinted in separate collections. Even if such editorial workflow has produced invaluable results for reconstructing lost authors, the concept of **textual fragment** is problematic: It includes different kinds of text reuse and implies a certain degree of originality, which is difficult to assess and represent because the original text from which the reuse derives is hidden by the **cover text**, i.e., by the intention of the quoting author and the characteristics of the preserving context (Most, 1997; Schepens, 2000; Berti, 2013).

Our data model defines taxonomies of text reuse for representing references to authors and works not as separate chunks of text but as contextualized annotations, expressing their nature of reuse of textual evidence. These annotations include not only the portion of text classifiable as a reuse, but also biographical and bibliographical data preserved in the source text.

Text reuse of fragmentary authors presents the challenge of documenting text aligned with no extant exemplar. Text reuse of extant works presents additional challenges of aligning as precisely as possible (but no more precisely than is possible) two or more extant passages of text that may differ in small ways or large. Our data model documents uniquely instances of text reuse and it is developed on the Canonical Text Services (CTS), which is a protocol for identifying and retrieving passages of text based on concise, machine-actionable canonical citation. It is founded on the assumption that a "text" can be modelled as "an ordered hierarchy of citation objects" (Smith and Weaver, 2009). CTS URNs can identify passages more grossly or more finely; they can identify a range of passages at various levels of specificity; by the addition of an indexed substring, a CTS URN can identify a particular string within a passage of text (Blackwell and Smith, 2012). CTS is one component of a larger digital library architecture, developed for the **Homer Multitext** project and called CITE (Collections, Indices, Texts, and Extensions): http://www.homermultitext.org/hmt-doc/cite/.

In order to produce citable analyses of text reuse in their context, we have been working with the *Deipnosophists* of Athenaeus of Naucratis, which is the account of a banquet where learned men quote authors and works of Greek literature concerning a wide range of topics related to dining and food. The *Deipnosophists* is significant because it is a very rich collection of many different kinds of text reuse of fragmentary authors and of extant works (Braund and Wilkins, 2000; Lenfant, 2007; Jacob, 2013).

Our data model specifies four subjects of analyses:

1. **Authors**: enumerate and identify authors reused by Athenaeus;

2. **Works**: enumerate and identify works reused by Athenaeus;

3. **Mentions**: catalog every mention of authors and works in the text of Athenaeus, including his vocabulary for identifying them. For example, Athenaeus may mention that a work by Archestratus of Syracuse was known by four different names (i.e., *Gastronomy*, *Life of Pleasure*, *Science of Dining*, or *Art of Cooking*); this would generate five entries in this list: one mention of Archestratus, and four mentions of the same work.

4. **Reuses**: uniquely identify instances of text-reuse in the text of Athenaeus.

A fifth analysis will also include the twenty-two learned men who take part in the banquet described by Athenaeus and who are actually the **characters** who quote and reuse a huge amount of authors and works.

We need seven records to produce citable analyses of the above mentioned subjects:

1. **Analysis Record URN**. Every documented instance of text reuse (authors, works, mentions, reuses) has a CITE URN uniquely identifying this instance in a CITE collection.

2. **Sequence Number**. The collection of instances of text reuse is an **ordered collection**; each item has a sequence number, reflecting the item's sequence in the text of Athenaeus. This value is programmatically generated by a CTS-aware script before publishing the collection.

3. **Analyzed Text**. A CTS URN defining, as precisely or imprecisely as necessary, the span of text in the *Deipnosophists* that is the subject of this analysis of text reuse. The scope of the **Analyzed Text** is determined by the nature of the text reuse. In the case of authors and works, this CTS URN identifies a passage in the *Deipnosophists* that serves to justify the inclusion in the respective list. When an author or a work is reused often, the passage should be a clear, unambiguous reference (e.g., "Homer says …").

4. **Reused Text**. While the **Analyzed Text** identifies a coherent and contiguous span of text, as it appears in the edition being analyzed, the **Reused Text** is a string identifying only the text being reused. The **Analyzed Text** provides context and a basis for alignment, while the **Reused Text** gives us the flexibility to call out non-contiguous text, to normalize text, or even to promote morphological forms determined by indirected statement to those appropriate for direct speech, without doing violence to our source-edition. The **Reused Text** record allows us to represent different intepretations of the same text reuse, especially in the case of non-verbatim quotations.

5. **Alignment URN**. This collection documents reuse of extant authors and works, for which we have extant editions with canonical citation. The **Alignment URN** is a CTS URN pointing to the quoted extant author (identified with a CtsGroupUrn) or to one specific edition of the reused work (identified with a CtsWorkUrn) that (a) justifies our claim of text reuse, and (b) is the basis for attaching a citation of a still extant work to this analysis.

6. **Analytical Edition URN**. The collected instances of text reuse of extant work in the *Deipnosophists* represent a new edition of these works, whose text-content is based on our analysis of our project's edition of Athenaeus. The **Analytical Edition URN** is a CTS URN to an **Athenaeus Edition** of these works; the citation-value is based on that of the **Alignment URN**; the text-content of this edition is the **Reused Text** in Athenaeus. The **Analytical Edition** gives us an orthogonal view of the text reuse of extant authors in Athenaeus.

7. **CITE Collection of Lost Works**. For text reuse of lost authors and works, there is no citation scheme, nor any inherent order to the text. For these, we produce a collection of text-reuse. This Collection can be cited by CITE URNs.

Initial work on documenting text reuse has been focused on references to Homer's *Iliad* in the *Deipnosophists* (data available at http://digitalathenaeus.github.io/). The aim is to extend our data model including the categorization of different kinds of text reuse and further concrete examples of references to fragmentary authors and extant works in the *Deipnosophists* of Athenaeus of Naucratis.

Bibliography

Berti, M. (2013). Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres. *Ancient Society* 43: 269–88. doi:10.1145/1555400.1555442.

Berti, M., Romanello, M., Babeu, A. and Crane, G. (2009). Collecting Fragmentary Authors in a Digital library. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM, pp. 259–62. doi:10.1145/1555400.1555442.

Blackwell, C. W. and Smith, D. N. (2012). Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture. In Muellner, L. (ed.), *Donum Natalicium Digitaliter Confectum Gregorio Nagy Septuagenario a Discipulis Collegis Familiaribus Oblatum*. Washington, DC: The Center for Hellenic Studies. http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=4846.

Braund, D. and Wilkins, J. (2000). *Athenaeus and His World. Reading Greek Culture in the Roman Empire*. Exeter: University of Exeter Press.

Jacob, C. (2013). *The Web of Athenaeus*. Center for Hellenic Studies: Harvard University Press.

Lenfant, D. (2007). *Athénée et les fragments d'historiens. Actes Du Colloque de Strasbourg (16-18 Juin 2005)*. Paris: De Boccard.

Most, G. W. (1997). *Collecting Fragments. Fragmente sammeln*. Göttingen: Vandenhoeck & Ruprecht.

Schepens, G. (2000). Probleme der Fragmentedition. (Fragmente der Griechischen Historiker). In Reitz, C. (ed.), *Vom Text Zum Buch*. St. Katharinen, pp. 1-29.

Smith, D. N. and Weaver, G. (2009). Applying Domain Knowledge from Structured Citation Formats to Text and Data Mining: Examples Using the CITE Architecture. *Text Mining Services* 129. http://katahdin.cs.dartmouth.edu/reports/TR2009-649.pdf

Two Centuries of Russian Roads – Diachronic Study of Polysemy in the Context of Cultural Change

Anastasia Bonch-Osmolovskaya
abonch@gmail.com
National Research Unversity Higher School of Economics
Moscow, Russian Federation

Preamble

The goal of the study is to show links between lexical and social diachronic change. The study is conducted in the culturomics framework (Michel et al., 2011). In contrast to the Big data approach the study promotes the idea of medium data, i.e. amount of data which allows both to make quantitative and qualitative analysis (Bonch-Osmolovskaya, 2015). The main characteristics of the medium data are:

- The reliability of sources, which metadata can be filtered manually
- The sufficiency of the data amount for reliable statistical measures
- The possibility of additional semantic mark-up

The research is based on the data from Russian National Corpus (ruscorpora.ru) (see Plungian, Sitchinava, 2003). The study pursues changes of context frequencies for the lexeme road in the period from 1800 till 2000, and correlates the observations with social and economic progress as well as change in conceptual language space

Choice of concept

Russia is a big country, so transportation has been traditionally a critical problem. The choice of the word *road* for culturomics study is based on our expectations of the concept's centrality for the economy, society and culture in Russia of the 19th–20th centuries. *Road* appears to be a productive sign in terms of semiotics of art (Tchepanskaya, 2003,), that's why I expected to collect numerous relevant contexts both in fiction and nonfiction. At the same time *road* in Russian has several meanings, the nature of its polysemy has been treated a lot in previous works (Arutiunova, 1999). We can distinguish three basic meanings which are contrasted by the position of Observer (Paducheva, 2006) – the one that percepts the *road*. The first meaning is *road* as a physical object, a line on the ground the observer sees while standing on it. It can be characterized by the quality of its surface or surrounding landscapes (i.e. dirty road). The second meaning is *road* as a vector, a line on a map, that connects two points (i.e central road). The observer operates in this case with the abstract idea of the road's topology. The third meaning is metonymical and it stands for the travel-event the Observer experiences while moving along the road (i.e. tedious road). Finally due to semiotic abundance *road* is frequently used in metaphorical sense (i.e. life path = "road of life"). At the same time, the first three meanings present the most important parameters that determine mobility of population: quality of roads, connectedness between localities and time and quality of journey. Therefore, it seems insufficient to track frequency change of *road* occurrences in the corpus in general, but it is important to distinguish how the frequency of different meanings has been changing.

Method

Different meanings of *road* can be captured by attributive constructions as adjectives usually refer to only one sense. The corpus has been divided into 7 time periods from 1800 to 2000. To make the sub-corpora comparable the 19th century has been divided into two periods of 50 years and the 20th century into five periods of 20 years. The contexts, containing constructions of adjective plus *road* has been extracted from every sub-corpus. The noisy entries has been removed, the data has been lemmatized and normalized as ipm. As a result, I obtained a database with 15000 constructions, containing more than 1500 unique adjectives.

On the next step, all the adjectives have been categorized by 20 semantic domains. The domains correlate with four basic meanings of *road* defined below but render more specific characteristics of different *road* parameters. The most frequent construction "zheleznaya doroga" (literary *metal road*, meaning railroad) has been selected in a separate category.

Then I applied hierarchical clusterization to the data of 20 categories, see Figure 1

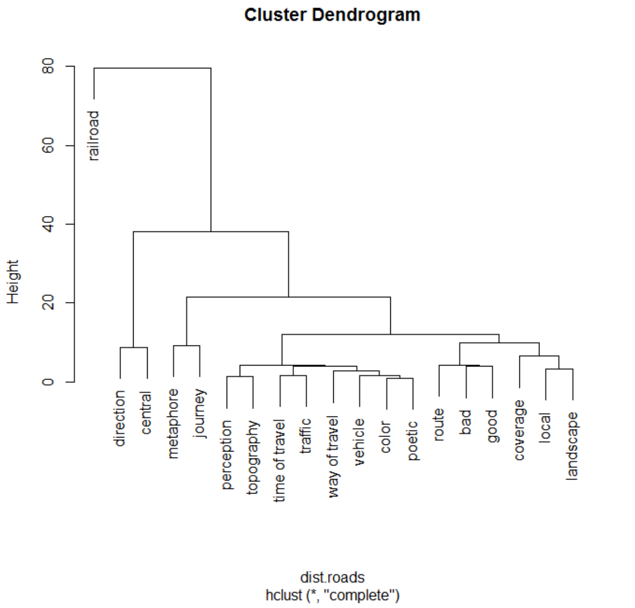
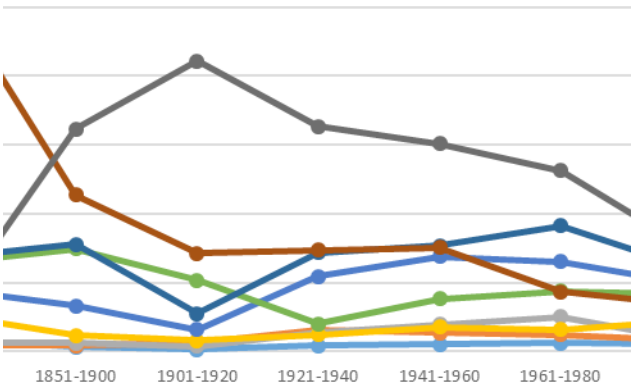


Figure 1: Clusterization of semantic categories for adjectives describing road

The data of the categories in one cluster has been summarized and then plotted on the graph (see Figure 2)



Analysis

The data allows plenty of research scenarios, comparing different domains, such as, for example:

- sources of domain contents and diachronic change of the distribution of sources (for example, fiction or nonfiction)
- widening or narrowing of the category through time (how many adjectives are in the category), as a well as persistence of the content to time and economical or social changes
- substitution of one adjective to another (for example, all the changes connected with replacement of horses to cars) and its time frame
- migration of an adjective from one adjective to another

In this abstract, I will focus on the most prominent changes of cluster graphs. As Fig 2 shows the railroad (RR) cluster and the direction and centrality(D and C) cluster are the most distinctive in their behavior. In the beginning of the 19th century, the existence of big central roads from one town to another completely determined mobility opportunities of Russian population. We see that more than 50% of all the occurrences of *road* are associated with D and C attributes (Warsaw road, big road – as a specific term of central road). In the 1851, the railroad between Moscow and St.Peterburg has been open and this fact nicely correlates with the crossing of the RR and D and C graph in the period of the 1850s. The intensive growth of the RR cluster in the second half of the 19th century reflects not only the growth of railroad communications in Russia but also great conceptual influence of the railroad innovation, which can be also traced in numerous literary pieces of this period such as Tolstoy's Anna Karenina or Dostoevsky's Idiot. The intriguing fact about RR cluster is its consistent fall in the 20th century that may of course correlate with developing automobile transportation. The sharp fall of

RR cluster in the 1960s corresponds to the growth of civil airlines; see Figure 3 that demonstrates quite the opposite trend for air transportations starting from the 60s

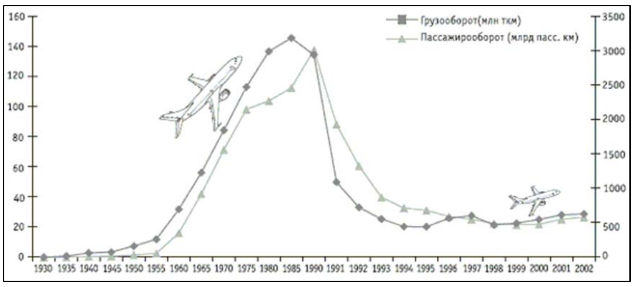


Figure 3: Graph of air transportation, line with triangles marking passenger traffic

The most important generalization that arises from the observations above is that in the 20th century the topological (vector) meaning of *road* is consistently fading while the reference to a road as a physical object on the contrary increases in frequency. In other words, while economic and industrial growth results in diversity of mobility means, *road* as a concept in lexical space has changed the balance of its meanings reducing the *connection* idea. At the same time the metonymic usage of *road* as a journey has been increased in the 20th century as a well as the metaphorical usage, the both categories are very similar in their data values so they have formed one common cluster. In 1960s, the *connection* idea is transferred from direct usages of D and C cluster to figurative usage of Journey and Metaphor cluster. This means that we can document the moment when the idea of *connection* is separated from the physical movement along the road. The tedious road now is sitting in the airport for many hours waiting for your flight.

Bibliography

Arutiunova, N. D. (1999). Put' po doroge I bezdorozhju [The way on the road and off the road]. In Arutiunova N.D., Shatunovskii I.B. (eds.), *Logicheskii analiz yazyka. Yazyki dinamicheskogo mira* [Logical analysis of language. The Languages of the Dynamic World]. Dubna, pp. 3-17.

Bonch-Osmolovskaya, A. A, (2015). Medium data method for cultural studies: the case of gender studies in Russian National Corpus, *Proceedings of Digital Humanities*, Sydney.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Aiden, E. L., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, **331**(6014): 176-82.

Paducheva, E. V. (2006). Nabludatel': tipologiya I vozmozhnye traktovki [The Observer: Typology and Possible Interpretations]. *Trudy mezhdynarodnoy konferentsii "Dialog-2006"*. [Proceedings of International Conference "Dialog-2006"]. Moscow, RSUH, pp. 403-13.

Plungian, V. A., Sitchinava, D. V. (2004). Natsionalny korpus russkogo jazyka:oput sozdaniya korpusov tekstov sovremennogo russkogo jazyka [Russian national corpus: experience of creating corpora for contemporary Russian]. In Beliaeva et al. (eds), *The proceedings of the international conference "Corpus Linguistics-2004"*, Saint-Petersburg, pp. 216-39.

Tchepanskaya, T. B. (2003). *Kultura dorogi v russkoi miforitual'noi traditsii XIX-XX vekov* [Road culture in Russian mythic and ritual tradition of the 19th-20th centuries]. Moscow, Indrik Publ.

Ancient Maya Writings as High-Dimensional Data: a Visualization Approach

Gulcan Can
gcan@idiap.ch
Idiap Research Institute and EPFL, Switzerland

Jean-Marc Odobez
odobez@idiap.ch
Idiap Research Institute and EPFL, Switzerland

Carlos Pallán Gayol
pallan.carlos@gmail.com
Abteilung für Altamerikanistik und Ethnologie, University of Bonn, Germany

Daniel Gatica-Perez
gatica@idiap.ch
Idiap Research Institute and EPFL, Switzerland

Introduction

The ancient Maya civilization flourished from around 2000 BC to 1600 AD and left a great amount of cultural heritage materials, in the shape of stone monument inscriptions, folded codex pages, or personal ceramic items. All these materials contain hieroglyphs (in short glyphs) written on them. The Maya writing system is visually complex (Fig. 1) and new glyphs are still being discovered. This brings the necessity of better digital preservation systems. Interpretation of a small amount of glyphs is still open to discussion due to both visual differences and semantic analysis. Some glyphs are damaged, or have many variations due to artistic reasons and the evolving nature of language.

Signs following ancient Mesoamerican representational conventions end up being classified according to their appearance, which leads to potential confusions as the iconic origin of many signs and their transformations through time are not well-understood. For instance, a sign thought to fall within the category of 'body-part' can later be proven to actually correspond to a vegetable element (a different semantic domain). Similarly, several signs classified as 'abstract', 'square' or 'round' could actually be pars-pro-toto representations of a larger whole.

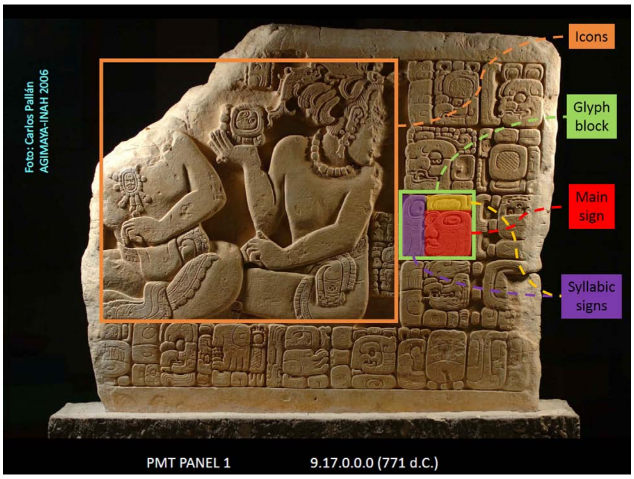


Figure 1. A stone inscription found in Pomona, Tabasco (Mexico), Panel 1 from 771 AD (Photograph by Carlos Pallán Gayol for AJIMAYA/INAH Project© 2006, Instituto Nacional de Antropología de Historia, Mexico)

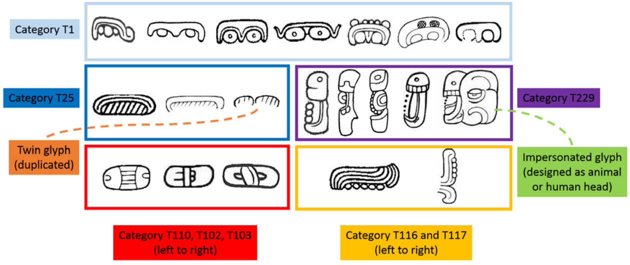

Category T1
Category T25
Category T229
Twin glyph (duplicated)
Impersonated glyph (designed as animal or human head)
Category T110, T102, T103 (left to right)
Category T116 and T117 (left to right)

Figure 2. Maya glyph samples from several categories (according to Thompson's catalog) that illustrate the within-class variety and between-class similarity

Fig. 2 illustrates the challenges to analyse Maya glyphs visually. Adding functionalities that take context (i.e., co-occurrence statistics, characteristics of the data) and part-whole relations (i.e., highlighting diagnostic parts) into account would bring guidance during decipherment tasks. The tools we envision are different from existing almanac-by-almanac visualization systems (Vail and Hernandez, 2013). They are also more engaging for users (i.e. visitors in museums), and offer promising perspectives for scholars.

This motivates the study of data visualization. In this paper, we built a prototype for visualization of glyphs based on visual features. We introduce (1) an approach to analyse Maya glyphs combining a state-of-the-art visual shape descriptor, and (2) a non-linear method to visualize high-dimensional data. For the first component, we use the histogram of orientation shape context (HOOSC) (Roman-Rangel et. al., 2011a; Roman-Rangel et. al., 2011b; Roman-Rangel et. al., 2013) which has similarities to other descriptors of the recognition literature (Belongie et. al., 2002; Dalal and Triggs, 2005; Lowe, 2004), but is adapted to shape analysis (Franken and van Gemert, 2013).

For the second component, we use the t-distributed Stochastic Neighbourhood Embedding (t-SNE) (Van der Maaten and Hinton, 2008), which is a dimensionality reduction method from the machine learning literature that has value for Digital Humanities (DH), as it can highlight the structure of high-dimensional data, i.e., multiple viewpoints among samples.

As analysis of DH data is often based on attributes like authorship, produced time, and place, observing these variations as smooth transitions with t-SNE becomes a relevant feature.

We show that the proposed methodology is useful to analyse the extent of spatial support used in the shape descriptor and to reveal new connections in the corpus through inspection of glyphs from stone monuments and glyph variants from catalogue sources. In particular, we hope that the presentation of our use of t-SNE can motivate further work in DH for other related problems.

Methodology

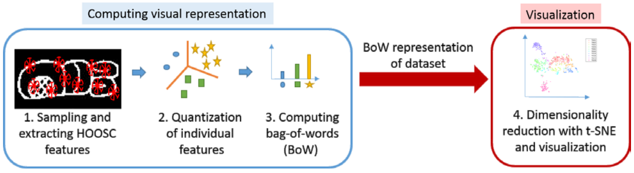

Computing visual representation
1. Sampling and extracting HOOSC features
2. Quantization of individual features
3. Computing bag-of-words (BoW)
BoW representation of dataset
Visualization
4. Dimensionality reduction with t-SNE and visualization

Figure 3. Overall flow for visualization with t-SNE

The analysis process is illustrated in Fig. 3. First, for each glyph, a standard visual bag-of-words representation (BoW) is computed from the HOOSC descriptors. Second, dimensionality reduction is performed on the BoW representation of a glyph collection to generate the visualization. The main steps are described below.

Datasets

We analyse our visualization pipeline on two individual Maya glyph datasets.

Monument data

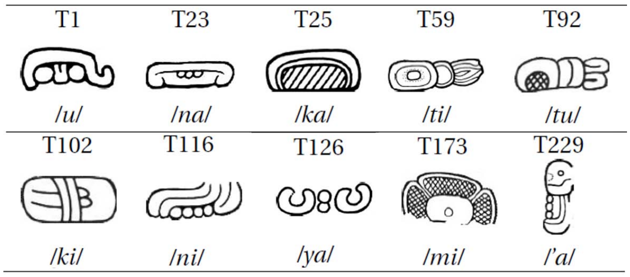


Figure 4. Sample glyph images, corresponding Thompson annotations, and syllabic values (sounds) of selected 10 classes from the syllabic monument glyph dataset

We use a subset (630 samples from 10 classes, Fig. 4) of hand-drawings (Roman-Rangel et. al., 2011), corresponding to syllabic glyphs inscribed in monuments. These samples are collected by archaeologists (as part of Mexico's AJIMAYA project) from stone inscriptions spread over four regions (Peten, Usumacinta, Motagua, and Yucatan). As an additional source, around 300 glyph samples are taken from existing catalogues (Thompson and Eric, 1962; Macri and Looper, 2003).

Thompson catalogue

Secondly, we use 1487 glyph variants cropped from the Thompson's catalogue. These variants belong to 814 categories and divided as main sign and prefix/suffix groups in the catalogue.

Visual feature representation

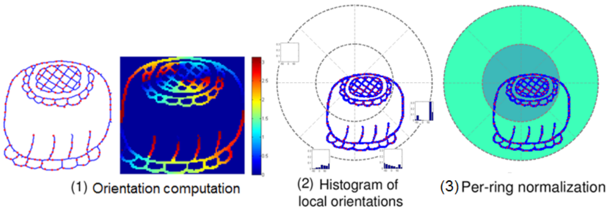

(1) Orientation computation
(2) Histogram of local orientations
(3) Per-ring normalization

Figure 5. HOOSC computation at a sample position of the shape

The HOOSC is a shape descriptor proposed in our research group for Maya glyphs (Roman-Rangel, 2011b). It is computed in two main steps (Fig. 5). First, the orientations of a set of sampled points are computed. Secondly, for a given sampled position, the histogram of local orientations are computed using a small number *Na* of angle bins forming a circular grid partition centred at each point. The HOOSC descriptor is obtained by concatenating all histograms, and applying per-ring normalization. Basic parameters are the spatial context *sc* defining the extent of the spatial partition; the number of rings *Nr*; and the number *Ns* of slices in a ring. With *Na* =8, *Nr* =2, *Ns*=8, HOOSC has 128 dimensions. We have used HOOSC for usual retrieval and categorization tasks (Hu et. al., 2015).

Dimensionality reduction: t-SNE

Proposed in (Hinton and Roweis, 2002), SNE is a non-linear dimensionality reduction method. It relates the Euclidean distances of samples in high-dimensional space to the conditional probability for each point selecting one of the neighbours. In t-SNE (Van der Maaten and Hinton, 2008), these distributions are modelled as heavy-tailed t-distributions. t-SNE aims to find for each data point, a lower-dimensional projection such that the conditional probabilities in the projected space are as close as possible to those of the original space (measured with KL divergence (Kullback and Leibler, 1951)).

In our application, first, we project the BoW representation to a 30-dimensional space using PCA, then applied t-SNE to these projections to get 2-dimension mapping. t-SNE keeps track of the local structure of the data as it optimizes the clusters globally.

Results and discussion

The full-scale visualization of the glyphs are available at https://www.idiap.ch/project/maaya/demos/t-sne.

Glyph monument corpus structure

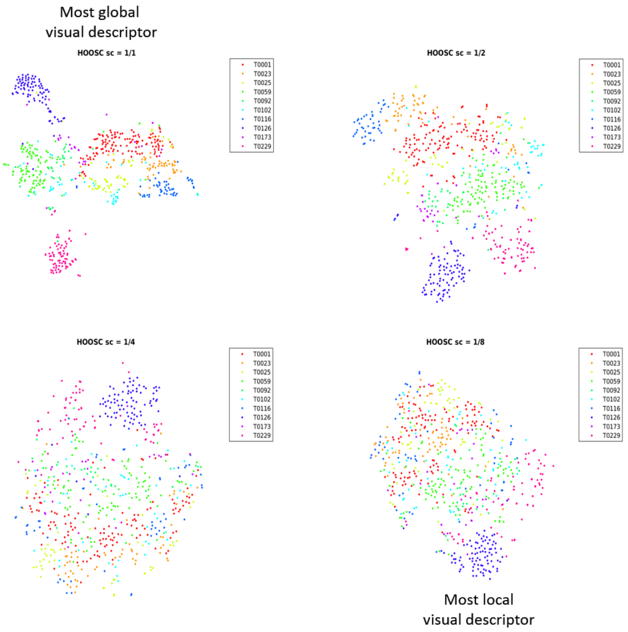


Figure 6. Monument data: t-SNE plots with visual representations obtained at four different spatial context levels

Fig. 6 shows the monument corpus. The region encoded in the visual descriptor varies from almost whole glyph (sc=1/1) to small local parts (sc=1/8). One question is how spatial context influences visualization of the representation. Regarding the visual clusters, with the most global representation (sc=1/1), our method extracts more distinct clusters, e.g. T229 and T126 in Fig. 7 (navy and magenta in Fig. 6 and 9). Please see Fig. 9 for roughly-coloured clusters of the glyphs. As the descriptor gets more local, the categories with common patterns mix up (Fig. 6). Yet, our method is able to capture meaningful common local parts and maps the samples based on these elements, i.e. parallel lines, hatches, and circles.

For Maya epigraphers in our team, a more neatly differentiated grouping of signs, e.g. obtained by HOOSC with sc=1/1 is preferable. However, work on the effects of parameter choice is required to obtain groupings that make more epigraphic sense. Clearer 'borderlines', less 'outliers,' and less 'intrusive' signs (e.g. T25 and T1) within each cluster would be desirable. Our results in this regard are preliminary, but they open promising research questions.

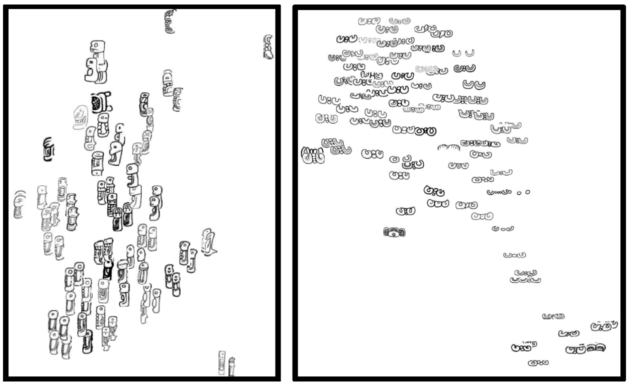


Figure 7. Monument data: Close-up of two clusters (T229 on the left and T126 on the right), corresponding to navy and magenta clusters in Fig. 6 with the most global HOOSC descriptor (sc=1/1)

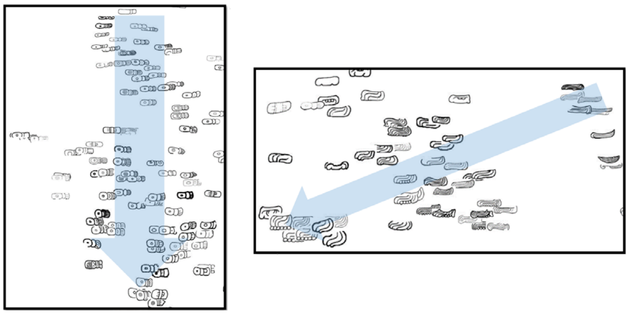


Figure 8. Monument data: Close-up of two clusters (T59 on the left and T116 on the right), which exhibit smooth transition between samples corresponding to place or temporal variations

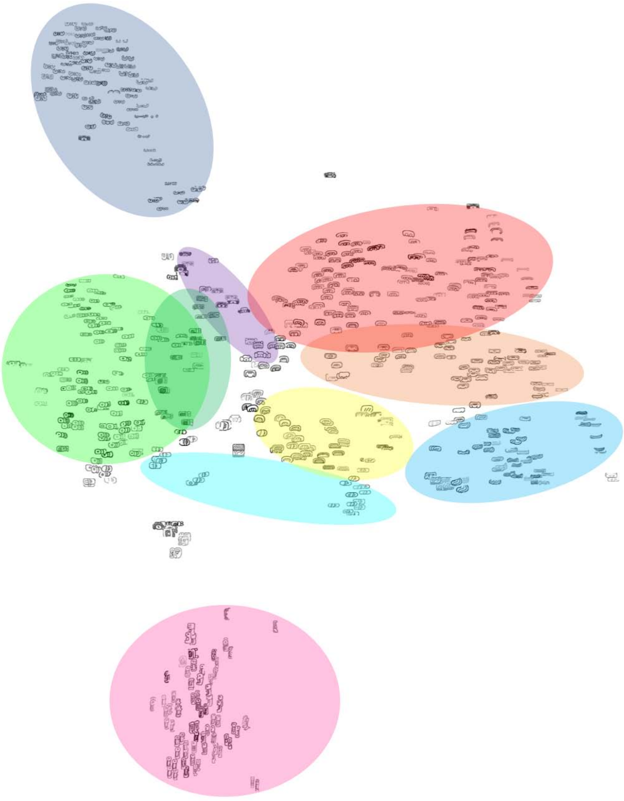


Figure 9. Monument data: Visualization of all class samples with the most global HOOSC descriptor (sc=1/1)

Another important epigraphic point is that we observe interesting visual transitions between samples of the categories. Fig. 8 shows examples from category T59 and T116, which illustrate a smooth dilation of samples in one direction. These kind of observations are interesting for archaeologists, since they might correspond to modification of the glyph signs over time or place.

Glyph variants from Thompson catalogue

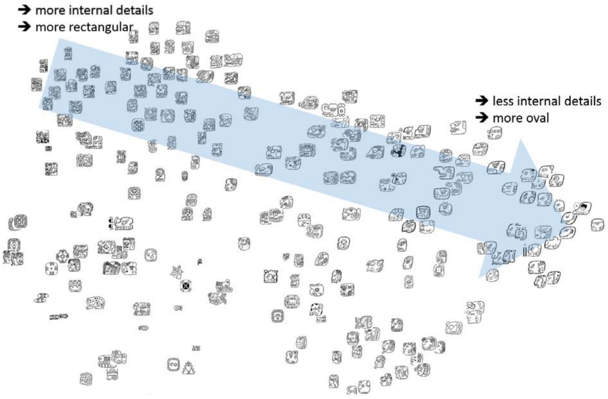


Figure 10. Catalogue data: A visual cluster of main signs from the Thompson's catalogue, with the most global HOOSC descriptor (sc=1/1). Many of them are impersonated main signs that corresponds to gods or animals. In this part of the visualization, the upper left part has more visually complex variants than the rightmost samples

From the visualization of glyph variants in Thompson's catalogue with the largest spatial context level (sc=1/1), we observe that visually similar categories are grouped together, while exhibiting smooth transitions. These transitions may correspond to some characteristics of the data. Fig. 10 shows a cluster of personified main signs in which degree of visual internal detail decreases in the indicated direction. We also observe separate visual clusters for hatched, horizontal and vertical glyphs.

Conclusion

Our goal in this study is to help DH scholars to visualize data collections not as isolated elements, but in context (visually and semantically). Even though early catalogues are built based on visual similarities, i.e., (Thompson and Eric, 1962) or (Zimmermann, 1956) relied on graphic cards to study similar patterns, the categorization methods were poorly understood and were not easy to reconfigure.

Furthermore, due to the limited knowledge at the time about semantics and sign variants, these catalogues turned out to be inaccurate or outdated. Similarly, Gardiner's list (Gardiner, 1957) is insufficient to elucidate sign variability in the 'Book of The Dead' (Budge, 1901).

With the proposed tool, however, considering details at different scales as semantic/diagnostic regions in the visualization can help archaeologists to discover semantic relations. In this way, overlapping notions such as 'colours', 'cardinal directions' and specific toponyms from earthly, heavenly or underworld realms can be studied in greater detail.

Finally, illustrating all variations with different visual focus in a fast and quantitative manner brings out the characteristics of signs. This also helps experts match samples from various sources (i.e. monuments, codices, and ceramic surfaces) to corpus data more efficiently; and trigger the decipherment of less frequent and damaged signs. Hence, our work is a step towards producing a more accurate and state-of-the-art sign catalogue.

Acknowledgements

This work was funded by Swiss National Science Foundation as part of the MAAYA project.

Bibliography

Belongie, S., Malik, J. and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(4): 509–22.

Budge, E. A. W. (1901). *The Book of the Dead: An English Translation of the Chapters, Hymns, Etc. of the Theban Recension, with Introduction, Notes, Etc.* (Books on Egypt and Chaldaea). Open Court Pub.

Dalal, N. and Triggs, B. (2005). *Histograms of Oriented Gradients for Human Detection. vol. 1. IEEE*, pp. 886–93.

Franken, M. and Gemert, J. C. van (2013). *Automatic Egyptian hieroglyph recognition by retrieving images as texts*, ACM Press, pp. 765–68.

Gardiner, A. H. (1957). *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs. 3d ed., rev*. Oxford: Griffith Institute, Ashmolean Museum.

Hinton, G. E. and Roweis, S. T. (2002). *Stochastic neighbor embedding*. pp. 833–40.

Hu, R., Can, G., Pallan Gayol, C., Krempel, G., Spotak, J., Vail, G., Marchand-Maillet, S., Odobez, J.-M. and Gatica-Perez, D. (2015). Multimedia Analysis and Access of Ancient Maya Epigraphy: Tools to support scholars on Maya hieroglyphics. *Signal Processing Magazine, IEEE*, **32**(4): 75–84.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**(1): 79–86.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60**(2): 91–110.

Maaten, L. Van der and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**(2579-2605): 85.

Macri, M. J. and Looper, M. G. (2003). *The New Catalog of Maya Hieroglyphs: The Classic Period Inscriptions*. University of Oklahoma Press. Vol. **1**.

Roman-Rangel, E., Odobez, J.-M. and Gatica-Perez, D. (2013). Evaluating shape descriptors for detection of maya hieroglyphs. *Pattern Recognition*. Springer, pp. 145–54.

Roman-Rangel, E., Pallan, C., Odobez, J.-M. and Gatica-Perez, D. (2011a). Analyzing ancient maya glyph collections with contextual shape descriptors. *International Journal of Computer Vision*, **94**(1): 101–17.

Roman-Rangel, E., Pallan Gayol, C., Odobez, J.-M. and Gatica-Perez, D. (2011b). *Searching the past: an improved shape descriptor to retrieve Maya hieroglyphs. ACM*, pp. 163–72.

Thompson, J. E. S. and Eric, S. (1962). *A Catalog of Maya Hieroglyphs*. University of Oklahoma Press Norman.

Vail, G. and Hernández, C. (2013). The Maya Codices Database, Version 4.1. *A Website and Database Available at: http://www.mayacodices.org/*.

Zimmermann, G. (1956). *Die Hieroglyphen Der Maya-Handschriften*. (Abhandlungen Aus Dem Gebiet Der Auslandskunde / Reihe B: Völkerkunde, Kulturgeschichte Und Sprachen). De Gruyter.

The Evolution of Virtual Harlem: Bringing the Jazz Age to Life

Bryan Wilson Carter
bryancarter@email.arizona.edu
University of Arizona, United States of America

Imagine being faced with the choice of remaining in your hometown, living in what you have known all of your life in harshly oppressive conditions, or leaving, along with a number of your family and neighbors, headed North or West toward a land of promise and opportunity. This was the basic choice that many African Americans had in increasing numbers after 1865 when the emancipation proclamation was signed and a flurry of legislation passed in an attempt to equal the playing field for recently emancipated African Americans. By the early 20th Century, the Great Negro Migration was well underway, as African Americans were being driven off the land by the violence and intimidation of the Ku Klux Klan, natural disasters, and increasing levels of legal and extra-legal oppression and inequality in all sectors of life. They were also being pulled to the North and West by the perceived promise of increased opportunities in many aspects of life, the receipt of letters and communication from friends and relatives who had already made the move, and the perception of less racism in the North and West.

By the end of World War I, The Great Migration had reached flood proportions with hundreds of thousands of African Americans moving to urban centers in the north such as Chicago, Il, Harlem, NY, the steel belt and the auto industry in the midwest, as well as out towards the great promise of the West. Those who left the south for the urban North knowingly or sometimes unknowingly became a part of the New Negro Movement , known later as The Harlem Renaissance, of the 1920s, and The Chicago Renaissance of the 1930s and 40s. Understanding the cultural, intellectual and political output of African Americans during this period is key to understanding American history and our contemporary society. But what was it like in the big city of the 20s? How did that change, if at all, from the big city of the 40s and 50s?Artistic production was huge during both these periods, how might similar issues of artistic production be discussed from within this virtual environment?

There are a variety of ways to introduce our current generation of students to this vibrant aspect of African American culture that encompass both traditional and avant garde approaches. This project builds on two existing projects to engage today's learners and teachers in ways not previously possible, and also intertwines modern technology and advanced visualization with traditional practices to create a learning experience like no other that is focused on *The Great Migration, and Beyond*. Additionally, this paper explores, through the discussion and demonstration of the aforementioned project, ideas related to digital preservation, lack of diversity in Digital Humanities and introduces Digital Africana Studies as an engaging pedagogy.

The Great Migration and Beyond represents the first time that two major virtual environments focused on African American history and culture will be connected through both technology and theme. We recognize that there will be a number of challenges as we create rich historic landscapes and experiences while simultaneously creating a non-linear, visually and technologically unified journey through time and place related to African American life and culture.

The technology enabling this virtual world development is based on the Unity 3D game engine. Unity 3D is a platform for creating virtual worlds and games that accepts 3D assets or models using Open Standards format, meaning that assets can originate in nearly any 3D modeling tool and those entities can then be imported into Unity. The platform then enables us to distribute the final project to the spectrum of media devices (e.g. PC, Tablet, Game Console, Mobile, and full VR environments such as the CAVE, CAVE2, or a Data Wall. Additionally, we plan to address how the environment can be experienced through the increasing number of peripheral devices such as the Oculus Rift, HTC Vive, Samsung Gear or even Google Cardboard.

Key to the verisimilitude and distinctive look/feel of the Harlem/Bronzeville environment is the use of motion capture-based animation of the characters that the visitor encounters. While exploring the environment, visitors will encounter scenes depicting significant historic events of the time and place or common everyday occurrences where, those occurrences actually become a part of the evolving experience, in essence, a living part of a Theater of the Surround.

The Starting Point

The Virtual Harlem Project is a representation of Harlem, New York, as it existed during the 1920s Jazz Age, in 3D space, where objects such as buildings, interiors, automobiles and more are constructed using 3D modeling applications. This project is one of the earliest full virtual reality environments created for use in the humanities and certainly one of the first for use in an African American literature course. Virtual Harlem was originally conceived by Dr. Bryan Carter as part of his graduate work at the University of Missouri-Columbia in 1996, and developed by the Advanced Technology Center there who assisted him in designing it as a collaborative Virtual Reality (VR) environment and learning simulation in which participants learned about and experienced the Harlem Renaissance to supplement real world courses about the subject. The Harlem Renaissance/New Negro Movement was a unique period in American history that occurred in the 1920's, just after the end of World War I.

It was a time when Langston Hughes, Eubie Blake, Marcus-Garvey, Zora Neale Hurston, Paul Robeson, and countless other African Americans made their indelible mark on the landscape of American and international culture. This is a time when African Americans made their first appearances on Broadway; chic supper clubs opened on Harlem streets; riotous rent parties kept economic realities at bay while the rich and famous, both white and black, attempted to outdo each other with elegant, integrated soirees (Lewis)., and African American artists and entertainers were the toast of European Cafe Society.

Time Machine: Bronzeville has been in-development as a multi-modal recreation and immersive experience of the vanished and historically significant Bronzeville section of Chicago's South Side, during The Chicago Renaissance period (1930–1950). Components of this project include a computer game treatment, online immersive web destination, and augmented reality gallery installation.

Through the use of digital 3D image creation and animation, game and web technologies, the visitor can explore the history, lore, and legends of Bronzeville during the defining events of the 20th century: The Great Migration, The Great Depression, The Chicago Renaissance, Jim Crow Segregation (American Apartheid), World War II, and the emergence of The Black Metropolis.

Intuitively navigating the avenues, alleys and interior spaces of Bronzeville, and interacting with its residents, the visitor discovers the genius, ingenuity, and invention in all the arts and humanities, from painters' studios and recital halls, juke joints and storefront churches, to lecture halls, theater stages and street corner soapboxes that distinguishes the vibrant and creative period of The Chicago Renaissance.

The Vision: An Expanded Harlem/Bronzeville World

The immersive cityscapes of *Virtual Harlem* and *Time Machine: Bronzeville* are historical simulations and interpretations of the vanished Harlem and Bronzeville of the 1920s, 1930s and 1940s, recreated from extensive research of the photo and graphic records, the African American press archives, personal memoirs, oral histories and statistics.

The richly detailed environments will be deeply embedded with informing and contextual media. The visitor's movements and interactions with animated characters and objects will trigger radio broadcasts, phonograph recordings, ambient soundscapes, and links to other resources. The visitor will be able to interact with and manipulate objects in the environment to access the archive of historical photos, print documents and media clips that illuminate the events, persons, and significance of the creative, cultural, social and commercial engines that were Harlem and Bronzeville in these periods. Extensive use is made of documentation gleaned from African American press, radio and film archives.

The projected Harlem/Bronzeville complex will enable the visitor to explore 3D simulations of the vanished Harlem and Bronzeville communities. With a click/touch, select locations can be viewed through decades of change, allowing the visitor to witness the evolution of the neighborhoods, and to compare the historical views with the contemporary cityscape. 3D animated scenes and tableaux depict historical events, persons and places significant in understanding this period and its arts movement. In-world interactive maps and guides aid the visitor in navigating the terrains and times.

Featured historical figures, locations, and significant events of Harlem and Bronzeville include: artists' profiles and portfolios for Langston Hughes, Richard Wright, Arna Bontemps, Margaret Walker, Gwendolyn Brooks, Inez Cunningham Stark, William Edouard Scott, Charles White, Archibald John Motley, Jr., Eldzier Cortor, William MacBride, Elizabeth Catlett, Gordon Parks, Horace Cayton, John Johnson, the Jones Brothers (Policy Kings), The Apollo Theater, the Cotton Club, Connie's Inn, The Theresa Hotel, The Dark Tower, The South Side Community Art Center, Parkway Community House, South Side Writers' Group, the American Negro Exposition, Louis Armstrong and Chicago Jazz, Thomas Dorsey, the "Father of Gospel Music", blues artists, Contralto Mahalia Jackson, Katherine Dunham and Ballets Negres, the Beaux Arts Ball, The Skyloft Players, The Bud Billiken Parade, Artists' and Models' Ball, Savoy Ballroom, Regal Theater, Rhumboogie, Club DeLisa, among many others.

Timeline, Maps, and Infographics

Through the Graphical User Interface Unity3d overlay (GUI), the visitor will be able to access a Timeline, maps and other infographics, presenting a multimedia chronology of The Great Migration, The Harlem Renaissance, and The Chicago Renaissance. Images and text, media clips and animations pop up as the cursor rolls over key points on the graph. Hot spots on the graph link to other components of the environment, enabling visitors to travel quickly around the space. The Timeline, maps and infographics also present contextual information about national and international events, as documented in the African American press and other media sources. This prototype GUI will also enable users to save information, take snapshots and submit suggestions for integration of their own research for approval or suggestions to improve the experience.

Community-Of-Interest

Another of our challenges will be to create a forum for a community-of-interest, encouraging inquiry and making connections, and a host site/server for live presentations, teaching/learning experiences, performances and discussions. We envision an expanding archive and repository for Humanities research and scholarship (Implementation Phase). The start-up phase will include the design and samples of supplemental materials and guides to aid educators, and researchers in making full use of the evolving technology components and capabilities of this multi-modal compendium of reference materials, image archive, bibliography, repository for and curation of visitor contributions, and library of, and portal to, essays and lectures by scholars.

The immersive Harlem/Bronzeville complex presents an unprecedented contribution to the historical representation and interpretation of African Americans, uniquely enabled by current and emerging technologies. The design of the content and interactive engagement will make the Harlem/Bronzeville destination attractive to a wide spectrum of visitors, with international reach.

Enhancing the Humanities

The evolution of video games, the excitement surrounding virtual reality with the launch of a number of VR headsets, the increased graphic and processing power of currently available personal computers and gaming consoles, the rise of low cost and accessible virtual environment and game development tools, and the focus on increased engagement in the classroom makes this a perfect time to consider ways that both these projects might focus on a much larger and unifying aspect of African American culture and transformative period of American history, *The Great Migration*. The nationwide observance of the centennial of The Great Migration begins in 2016, and our project is intended to contribute to this multi-year observance.

The Larger Discussions

This project, in addition to it representing an innovative use of emerging technologies to teach Africana content, also introduces three parallel discussions; digital preservation, lack of diversity in Digital Humanities and the pedagogy of Digital Africana Studies.

Lack of Diversity in Digital Humanities

So what exactly is the Virtual Harlem/Bronzeville project? Since its inception, the Virtual Harlem Project has been called a variety of things to include a virtual learning platform, a collaborative learning network and a digital humanities project. Although it is most likely one of the oldest VR environments focused on African American life and culture, more specifically, that of the 1920s Jazz Age/ Harlem Renaissance, until recently, Virtual Harlem has rarely been discussed as an example of diversity in Digital Humanities. This is rather odd given that the project is particularly focused on African American life and culture, that it was conceived by an African American scholar and that it was initially intended to be used in an African American literature course. Yet, discussions of Virtual Harlem tend to center on the project as a good example of advanced visualization, technology in the classroom, or even a digital humanities project, without much mention of it having an emphasis on diversity. There are a variety of reasons why this may be the case. One is that there are so few VR projects that deal with diverse topics that those that are tend to be discussed with regards to the technologies used, not necessarily the nature or content of the project. Virtual Harlem also represents the use of interactive technologies where students are encouraged to become active contributors to a much larger landscape that is accessible by other members of the class and by limited numbers who are not enrolled. Contributions to the environment are done in non-traditional ways, to include performance, 3D body scanning and motion capture. It is just this sort of non-traditional teaching, learning and use of advanced technologies that some, particularly within the Africana Studies scholarly community, sometimes tend to view with a level of scepticism. Possible reasons for this are complex and may in part be related to how traditionally, scholarly respect for the discipline has been connected to knowledge and experience of its scholars. When the focus is shifted towards the technology used, there may be a fear of losing part of that respect. Tara McPherson suggests that "politically committed academics ...engage technology and its production not simply as an object of our scorn, critique or fascination but as a productive and generative space that is always emergent and never fully determined" (McPherson 155). Virtual Harlem represents one of the myriad of technologies that are and can be used in the humanities to address the learning styles of this generation of students, to encourage students to make use of the variety of tools they have at their disposal to express their understanding of humanities content, and to help us all deal with increasing amounts of data and information directly and indirectly related to the humanities. There are, however, a number of exciting projects by a diverse set of scholars that are flying under the radar simply because these scholars tend not to publish their work in digital humanities journals, many are dealing with tenure and promotion and focusing their publishing efforts on that which those evaluating their dossiers are more familiar instead of pushing projects that some fear may call their scholarship into question simply because evaluating digital humanities projects is not always fairly done in every discipline. These scholars are caught in a catch 22. Their projects are an amazing example of the use of technology in the classroom, yet evaluators have a fundamental lack of understanding of what digital humanities is or how to deal with it for tenure and promotion. So underrepresented scholars typically have to publish in traditional venues while working on their digital humanities projects as side projects until tenure and promotion are earned. Funding for digital humanities projects dealing with diverse topics is also difficult. Statistics are difficult to come by, but judging from the articles published in the most difficult popular digital humanities journals and books, there are very few pieces published that document projects dealing with diverse topics, created by underrepresented scholars. That is a problem. So what can be done to diversify the field? What can be done to encourage young minority scholars to contribute their ideas to the growing body of digital humanities scholarship in an effort to strengthen the field? Answers to these questions are not easy, nor will they be addressed in the very near future. However, there are efforts being made to introduce underrepresented scholars to digital tools and projects that, in time, may filter into the classroom and eventually into scholarship of these groups. These activities include workshops funded by the National Endowment for the Humanities and the Mellon Foundation along with pedagogies being developed that inherently incorporate technology as a part of the teaching and learning process.

Digital Africana Studies

Digital Africana Studies closely parallels Digital Humanities in that it encourages scholars to use a variety of technologies to teach and research within their fields and seek connections to the outside world in an effort to understand the human condition. Furthermore, Digital Africana Studies, also encourages students to use a variety of digital tools to express their understanding of course content and find relationships between what they are learning, their lives and the world around them. Digital Africana Studies is a direct outgrowth of Afrofuturism, which is a theory that explores how people of African descent are represented in conversations of the future, whether that future be depicted through science fiction, the entertainment industry, popular culture or education, as well as a way of looking at the world, a canopy for thinking about black diasporic artistic production, a way of considering the presence of people of African descent within past, present and future Western society, as well as an epistemology that is thinking about the future, the subject position of black people, and about how that is both alienating and about alienation. So when I am asked, "what *is* Virtual Harlem?", I find it simultaneously strange, curious and sometimes insulting that some would try to categorize, label and subsequently only *see* it through a relatively narrow lens, thus missing the larger, experiential aspect of the project. Digital Africana Studies is the pedagogical and practical application of Afrofuturism in that it seeks to explore how advanced technologies may be used in the classroom to support a more experiential environment, one that creates memorable encounters with the culture and content of the course.

Bibliography

McPherson, T. (2012). Why are the Digital Humanities so White? Or Thinking the Histories of Race and Computation. In Gold, M.K. (ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 155.

Jonathan Edwards and Thomas Foxcroft: In Pursuit of Stylometric Traces of the Editor

Michał Choiński
michalchoinski@gmail.com
Jagiellonian University, Krakow, Poland

Jan Rybicki
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland

Jonathan Edwards (1703-1758) is generally considered the most eminent and versatile thinker in early American history. His impact on the shaping of the theological thought and the preaching tradition of the colonial period was profound and long-lasting. Today he remains one of the best studied figures of the American past and different elements of his impressive output are continually reprinted by both academic and commercial publishing houses. Over his life Edwards authored more than a thousand sermons, hundreds of letters and a number of theological treatises. The Jonathan Edwards Research Center at Yale University edited and published most of these texts in their complete form as *The Works of Jonathan Edwards*, led by Harry Stout as general editor and Kenneth Minkema as executive editor. The series of almost thirty volumes is described by Phillip Gura, a former editor of *Early American Literature,* as the "most important editorial project in American cultural history in the past 50 years" (2004, 149). Edwards' life is so well documented that there are hardly any stones unturned in the life of the Northampton divine. Especially, his relevance for the events of the Great Awakening, a powerful social-religious movement of colonial America, underwent close scrutiny and the most notorious sermon of America, "Sinners in the Hands of an Angry God" which he authored, has been the studied linguistically, rhetorically and stylistically.

Like most people of his age, Edwards was a diligent diary-keeper and an avid letter-writer. His private texts offer a comprehensive insight into his daily struggles and ambitions – in consequence, the very writing and publishing process of his texts is relatively well documented. Yet, surprisingly – in spite of such extensive research conducted upon Edwards – the relationship between him and Thomas Foxcroft (1697-1769), his editor and literary agent has not been extensively studied.

Foxcroft was a minister at First Church in Boston, Massachusetts and Jonathan Edwards's ally in the pro-revival debate. Their collaboration began most probably in 1849; Edwards had great trust in his erudition and skill to carry out the authorial intent expressed point-by-point in his commentaries to the suggestions of corrections. Foxcroft sometimes included Edwards's correction verbatim, exactly as indicated by the author, at other times, he paraphrased them, while preserving the author's thought. Edwards entrusted Foxcroft with the editing, the correction and the publication process – as he writes in a letter sent from Stockbridge – a small mission he was sent to after the dismissal from his own parish of Northampton: "I should be glad that you would endeavor that this book may be printed in a pretty good paper and character, and may be printed correctly, and that particular care may be taken that the printer don't skip over a whole line as they sometimes do. And if the bookseller can be agreed with to let me have a number for the copy, it would be pleasing". (30 June, 1752). Edwards also consulted Foxcroft about the correctness of his interpretation of other authors: "(…) it is very difficult, and almost impossible, for another to enter into all the views of a writer, or to know everything he has in view in all that he says; and therefore a little variation of sentiment, may much thwart and disappoint his design, insensibly to another. But this I should take as a very friendly part and much desire, that if you observe, that in any instances I have mistaken Mr. Williams' meaning, and misrepresented him, or in any respect injured him (…)" (30 June, 1752). The extent to which the style of the editor (whose idiosyncratic style can be described on the basis of numerous publications he himself authored) permeated the author's writings in this case has not been determined. The influence of Foxcroft's thought and style in Edwards's writings seems to be potentially very strong and demands close investigation and the stylometric approach seems a most fitting tool to be employed for such a study.

The analysis was performed with two quantitative methods: frequencies of most frequent words were compared between the texts using the Delta procedure (Burrows 2002); then, an analogous procedure (this time using Support Vector Machines) was used to look for traces of the editor's signal in consecutive segments of several treatises by Edwards ("rolling.classify," Eder 2015a). The analyses were performed with *stylo* (Eder et al. 2013), a package for R, the statistical programming environment (R Core Team 2014), postprocessed with Gephi network analysis software (Bastian et al. 2009).

A general view of stylometric similarities and differences between the writings of Edwards and Foxcroft is presented in the network diagram in Fig. 1. It shows, above all, a good separation of the signal of the two preachers, especially when Edwards's spiritual texts; sermons, treatises and Biblical comments are concerned, these, in turn, exhibit a degree of separation by subgenre – as opposed to Foxcroft's generally more uniform stylometry.

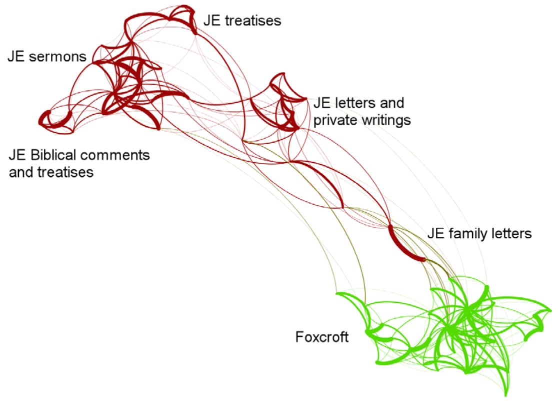


Figure 1. Network analysis of texts by Edwards (red) and Foxcroft (green).

In the more detailed search for the editor's signal with the "rolling.classify" method, longer texts by Edwards, i.e. his treatises, were compared against his own signal averaged over the rest of his *oeuvre* and against that of Foxcroft, bearing in mind the suggested caesura of 1749. Sure enough, consecutive segments of Edwards's works written before that date exhibited no traces of the editor (as exemplified by Fig. 2), and then surfaced in a series of works (as visible in Fig. 3). Interestingly, the editor's signal disappears again in 1758, the year of Edwards's (not Foxcroft's) death.

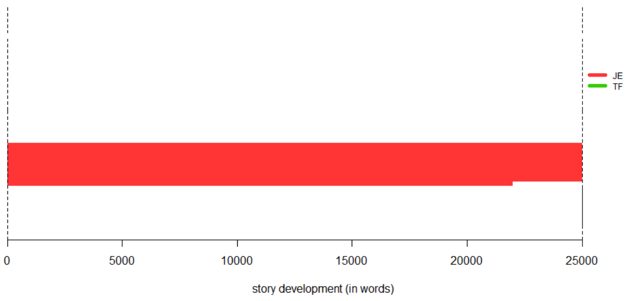


Figure 2. Consecutive segments of Edwards's *Mind* (1723); throughout the work, Edward's signal (red) dominates over the (absent) signal of Foxcroft.

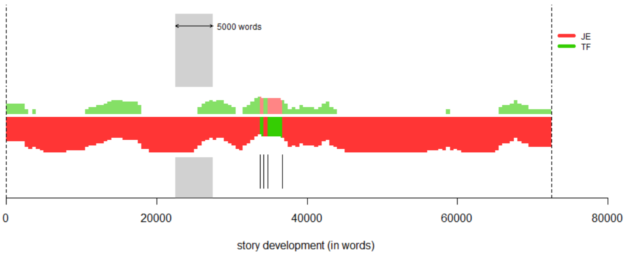


Figure 3. Consecutive segments of Edwards's Humble Inquiry (1749); in many other fragments, dominated by Edwards (red), Foxcroft's impact is still visible. The lower band shows the strongest signal; the upper, the second strongest.

These results have two significant consequences. The first is that we have now produced a quantitative confirmation of the extent of collaboration between two major colonial authors. But the fact that the quantitative agrees so well with the qualitative (or historical) evidence also shows that editorial traces can indeed be found with stylometry, perhaps to a greater degree than we might have anticipated.

Acknowledgements

This research is part of project K/PBO/000331, supported by Poland's National Science Center.

Bibliography

Bastian M., Heymann S. and Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

Burrows, J. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**: 267-87.

Eder, M., (2015a). Rolling stylometry. *Digital Scholarship in the Humanities*, **30**, first published online 7 April 2015, doi: 10.1093/llc/gqv010.

Eder, M. (2015b). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, **30**, first published online 3 December 2015, doi: 10.1093/llc/fqv061.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference abstracts*, University of Nebraska-Lincoln, pp. 487-89.

Gura, Philip F. (2004). Jonathan Edwards in American Literature, *Early American Literature* **39**(1): 147-166.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, **9**(6): e98679. doi:10.1371/journal.pone.0098679.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien, http://www.R-project.org/.

What's in a Topic ModelI fondamenti del text mining negli studi letterari

Fabio Ciotti
fabio.ciotti@uniroma2.it
Università di Roma Tor Vergata, Italy

Introduzione

Nel 1975 William A. Wood, ricercatore alla Bolt Beranek and Newman (una delle culle di Internet), pubblicò un articolo dal titolo "What's in a link" . Il lavoro, che determinò una svolta nella storia dell'Intelligenza Artificiale e in particolare della Knowledge Representation, consisteva in una serrata e argomentata critica della nozione di rete semantica (in tutte le sue varie accezioni, dal primo modello proposto da Ross Quillian alla più avanzata nozione di *conceptual dependencies* di Roger Shank), il concetto centrale di gran parte delle teorie di semantica computazionale. Il problema secondo Wood era che quella nozione non era fondata teoricamente (Wood, 1975): "there is currently no 'theory' of semantic networks. The notion of semantic networks is for the most part an attractive notion which has yet to be proven. Even the question of what networks have to do with semantics is one which takes some answering."

Nel suo titolo questo paper allude, immodestamente, al lavoro di Wood poiché si propone di affrontare, con lo stesso atteggiamento critico, i fondamenti teorici di una nozione e di un metodo che sono oggi molto diffusi negli studi letterari computazionali: quella di topic modeling, ovvero l'individuazione statistico/probabilistica dei cluster lessicali che caratterizzano un insieme di testi, e l'analisi delle loro distribuzioni (Underwood, 2012a; Blei, 2013). Occorre chiarire che quando parliamo di fondamenti teorici ci riferiamo al ruolo che tale nozione può giocare nel contesto di una teoria del testo e di una metodologia della critica letteraria, e non ai suoi aspetti puramente matematici, che sono ovviamente saldamente basati sulla statistica e sulla teoria della probabilità bayesiana (Blei, 2012).

Il distant reading e l'interpretazione

Come noto le tecniche di topic modeling rientrano nell'insieme più generale di metodi e approcci all'analisi computazionale dei testi letterari che Mario Moretti, con una formula di notevole successo coniata in antinomia con il metodo critico tradizionale, ha definito "distant reading" (Moretti 2013b: 48):

> what we really need is a little pact with the devil: we know how to read texts, now let's learn how not to read them. Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes— or genres and systems.

Tra i numerosi ricercatori che, con interesse se non addirittura con entusiasmo, hanno adottato questi metodi si nota talvolta una specie di atteggiamento riduzionista secondo cui i fenomeni letterari possono essere ridotti senza residui a fenomeni quantitativi misurabili e analizzabili in virtù di metodi puramente numerici e statistici (Jockers, 2013):

> Close reading is not only impractical as a means of evidence gathering in the digital library, but big data render it totally inappropriate as a method of studying literary history [...] massive digital corpora offer us unprecedented access to the literary record and invite, even demand, a new type of evidence gathering and meaning making.

A questo atteggiamento ovviamente si oppone una posizione che rivendica la radicale irriducibilità della letterarietà e/o dell'ermeneutica letteraria, la loro natura qualitativa. La tesi che intendo difendere in questo paper è la seguente: un metodo quantitativo o computazionale riveste interesse come strumento critico letterario nella misura in cui fornisce dati osservativi che possono essere correlati a termini o nozioni teoriche rilevanti per la teoria e la critica letteraria. In particolare, ritengo che le teorie semiotiche del testo letterario rappresentino una fonte di modelli e teorie di enorme utilità in ambito computazionale.

L'interpretazione di un testo letterario consiste nella creazione ed elaborazione di una serie di nozioni intenzionali adottate per spiegarne il funzionamento: nozioni come quella di storia, personaggio, eroe, autore e lettore, descrizione. Possiamo dire che la scuola semiotica nella teoria e nella critica letteraria ha adottato verso queste nozioni la stessa strategia che la corrente funzionalista nella teoria della mente ha avuto verso i termini della psicologia del senso comune: ne ha fornito una spiegazione in termini di concetti (come quella di attante, funzione narrativa, fabula, intreccio, isotopia) e modelli formali più rigorosi. Questi concetti non negano la natura intenzionale dell'interpretazione, ma non la intendono come un fenomeno irriducibile, bensì ne tentano una spiegazione al livello del progetto, per usare i termini del filosofo Daniel

Dennet (Dennet, 1990). Lo stesso Eco ha descritto questo doppio livello esplicativo introducendo la distinzione tra interpretazione semantica e interpretazione critico/semiotica (Eco, 1990: 29):

> L'interpretazione semantica o semiosica è il risultato del processo per cui il destinatario, di fronte alla manifestazione lineare del testo, la riempie del significato. L'interpretazione critica o semiotica è invece quella per cui si cerca di spiegare per quali ragioni strutturali il testo possa produrre quelle (o altre alternative) interpretazioni semantiche.

D'altronde è lo stesso Moretti (che in questo mostra una accortezza teorica e una conoscenza del campo letterario non riscontrabile in altri autori) a suggerire questa via in uno dei suoi migliori saggi teorici adottando la nozione epistemologica di "operazionalizzazione" come processo di traduzione di un termine teorico in una procedura sperimentale (Moretti, 2013a):

> Operationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts.

Topic model e loro possibili interpretazioni

Accettando il suggerimento metodologico di Moretti, quale nozione teorica della teoria letteraria viene operazionalizzata dal concetto statistico di *topic model*? O per usare un linguaggio più caro ai praticanti delle tecniche di *text mining*, di quale fenomeno letterario funge da "proxy" un topic model?

Dal punto di vista tecnico il *topic modeling* è una tecnica di text mining non supervisionata. Esistono diversi algoritmi di topic modeling, ma ad oggi il più diffuso è quello noto come Latent Dirichlet Allocation (LDA), fondato su un approccio probabilistico bayesiano. In modo intuitivo possiamo dire che alla base di LDA vi è un semplicistico modello generativo del testo: quando un autore scrive un testo in prima battuta sceglie l'insieme degli argomenti (topic) di cui vuole parlare e poi determina la proporzione con cui ciascun argomento sarà presente. Ammettiamo ora che ogni possibile topic possa essere caratterizzato come un insieme di parole con una data distribuzione: una specie di sacchetto di parole dove le parole possono essere ripetute in ragione diversa a seconda delle loro rilevanza rispetto all'argomento. Il nostro autore dunque potrà "pescare" in modo casuale dai vari sacchetti che corrispondo agli argomenti di cui intende scrivere ed estrarre da ciascuno un numero di parole proporzionale al peso che intende assegnare a ciascun argomento. Alla fine non dovrà fa altro che mettere in sequenza il suo mucchietto di parole ed ecco che avrà ottenuto il suo testo, in cui ovviamente le parole avranno una distribuzione determinata dalla rilevanza degli argomenti/sacchetti da cui sono state estratte. In termini tecnici si dice che in LDA un testo è una distribuzione di probabilità su un insieme di topic e un topic una distribuzione di probabilità su un insieme di parole. La cosa interessante di questo semplice metodo modello generativo è che può essere invertito: otteniamo in questo modo un algoritmo che è in grado di estrapolare i topic prevalenti in un insieme di documenti, ovvero la lista delle parole che co-occorrono con frequenza notevole e la loro distribuzione di probabilità.

I problemi tecnici dell'applicazione immediata di questo metodo (come di altri simili) in ambito letterario, sono diversi (Sculley and Pasanek, 2008). Ma qui ci interessa soprattutto il fatto che la nozione di topic non ha un chiaro statuto in ambito letterario: che cosa è quella lista di parole che costituisce un topic? E che cosa è la lista di topic (il topic model) nel suo insieme?

Le risposte possibili a questo quesito sono diverse: Ted Underwood ha proposto di interpretare i topic prodotti dal algoritmi come LDA come "discorsi", ovvero "kinds of language that tend to occur in the same discursive contexts" (2012a). Ma questa proposta da un lato richiede che lo statuto del concetto di discorso sia meglio definito; dall'altro nella formulazione di Underwood sembra dare adito a circolarità, poiché è proprio la natura e la funzione letteraria di quel particolare insieme di parole che va spiegata. Altri possibili candidati come correlati teorici della nozione di topic sono le nozioni di tema e motivo, che hanno avuto una lunga storia teorica nella teoria letteraria del secolo scorso; ma anche in questo le varie accezioni che possiamo assegnare ai concetti di tema e motivo paiono avere poco a che fare con i risultati di un algoritmo come LDA. Temi e motivi sono entità di contenuto che possono manifestarsi linguisticamente in specifici sintagmi o enunciati, ma che possono altresì essere manifestate da vaste porzioni di testo (a limite un testo nella sua interezza) senza avere nessun correlato linguistico immediato (Segre, 1985) . Si identificano in quanto temi e non generici contenuti concettuali in virtù della loro natura di stereotipi culturali che risiedono nella memoria culturale collettiva a cui autori e lettori attingono, pur mutandone nel tempo e nello spazio i valori semantici connotativi. Altrettanto problematica la interpretazione di un topic model come *isotopia* nel senso definito da Greimas (1985), intesa come classe paradigmatica di caratteristiche testuali discorsive che presentano una omogeneità semantica: infatti l'isotopia non è un fenomeno lessicale e soprattutto è il prodotto della cooperazione interpretativa del lettore, come la nozione di *topic discorsivo* adottata da Eco in nel suo *Lector in Fabula* (1979).

Conclusion

La discussione delle possibili interpretazioni semiotico letterarie della nozione di topic modeling e la constatazione della difficoltà teoriche che esse presentano ci porta ad affermare che in effetti non è possibile trovare un unico e soddisfacente correlato teorico- letterario dei risultati di questi metodi di analisi quantitativa. La conseguenza di questa difficoltà nella definizione teorica, ovviamente, non deve essere il rifiuto di queste tecniche di analisi come metodi utili alla conoscenza dei testi letterari. Occorre tuttavia essere consapevoli che di volta in volta il ricercatore dovrà individuare, sulla base dei testi che ha sottoposto ad analisi, quale siano i fenomeni letterari che i risultati intendono spiegare. Ne consegue inoltre la fallacia dell'idea di adottare nella ricerca letteraria le strategie di analisi esplorativa tipiche dei metodi della *Big Data analytics*. La natura intenzionale degli oggetti letterari rende impossibile individuare fenomeni rilevanti senza avere un modello o una ipotesi critica che orienti e regoli l'indagine.

Bibliography

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84.

Blei, D. (2013). Topic modeling and digital humanities. *Journal of Digital Humanities*, **2**(1).

Dennett, D. C. (1990). The Interpretation of Texts, People and Other Artifacts. *Philosophy and Phenomenological Research*, **50**(S): 177-94.

Eco, U. (1979). *Lector in fabula: la cooperazione interpretativa nei testi narrativi*. Milano: Bompiani.

Eco, U. (1990). *I limiti dell'interpretazione*. Milano: Bompiani.

Greimas, A. J. (1985). *Del senso 2: narrativa, modalità, passioni*. Milano: Bompiani.

Moretti, F. (2013a). Operationalizing: Or, the Function of Measurement in Literary Theory. *New Left Review*, **84**: 103-19.

Moretti, F. (2013b). *Distant Reading*. London: Verso.

Sculley, D. and Pasanek B. M. (2008). Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities. *Literary and Linguistic Computing*, **23**(4): 409–24.

Segre, C. (1985). Tema/motivo. In Segre, C. (ed.), *Avviamento all'analisi del testo letterario*. Torino: Einaudi, pp. 331-56.

Underwood, T. (2012a). Topic modeling made just simple enough. *The Stone and the Shell*. https://tedunderwood.wordpress.com/2012/04/07/topic-modeling-made-just-simple-enough/.

Underwood, T. (2012b). What kinds of "topics" does topic modeling actually produce?. *The Stone and the Shell*. http://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/.

Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In Bobrow D. G. ad Collins A. (eds.), *Representation and understanding: Studies in cognitive science*. New York: Academic.

An OWL 2 Formal Ontology for the Text Encoding Initiative

Fabio Ciotti
fabio.ciotti@uniroma2.it
Università di Roma Tor Vergata, Italy

Silvio Peroni
silvio.peroni@unibo.it
Università di Bologna, Italy

Francesca Tomasi
francesca.tomasi@unibo.it
Università di Bologna, Italy

Fabio Vitali
fabio@cs.unibo.it
Università di Bologna, Italy

Introduction

This paper presents the results of an effort that our research team has done in order to develop an OWL 2 ontology to formally define the semantics of the Text Encoding Initiative markup language. The preliminary steps of this research project have already been presented at the TEI Conference in 2014 and 2015 (Ciotti and Tomasi 2014, Ciotti et al., 2015). We believe that our work has reached a satisfactory level of development, both on the theoretical side and in the practical implementation.

Why an ontology for TEI

The reasons to have a formal and machine-readable semantics for TEI are manifold. In the first place we can set forth a list of pragmatic and technical benefits that have been already pointed out in many previous works dedicate to this topic, that dates back to the mid-90s (Di Iorio, Peroni and Vitali, 2009; Ciotti and Tomasi, 2014). Here is a brief summary of those arguments:

- enabling parsers to perform both syntactic and semantic validation of document markup;
- inferring facts from documents automatically by means of inference systems and reasoners;
- simplifying the federation, conversion and translation of documents marked up with different markup vocabularies;
- allowing users to query upon the structure of the document considering its semantics.

The advantages envisioned in this list are not specific to the TEI or aim to facilitate the relationships between different markup languages; but some of the issues have special relevance for TEI and for the usage of TEI inside its reference community.

Take for instance the query issue: we all know that there are many ways of expressing one and the same textual feature in TEI markup, so that it is very difficult to query heterogeneous TEI corpora and text archives. Having a set of ontological definitions of the conceptual level behind markup, that is, a set of shared formal definitions of the textual features to which any single encoding project could bind idiosyncratic markup usage, could help solve this problem. The same argument could be made for a far more adequate management of interoperability of TEI text collections between different repositories or applications.

But we believe there is also a deeper theoretical and foundational advantage in the idea of an ontological semantic model for TEI. It is a commonly acknowledged notion that the very core of digital methods application in humanities research is the notion of model/modeling. The pair of terms "model/modeling" is deplorably understood in many different ways in the community. We think that, as far as we are using Turing machine like device for computation, the only workable notion of modeling is a formal one: model we should be interested in are formal models. Where formalization is to be understood as a series of semiotic processes that generates an algorithmically computable representation of one (or more) phenomenon/object.

It is widely recognized that the TEI is not only a markup facility but first and foremost a conceptual model of textuality. In fact, the Guidelines (TEI Consortium, 2015, chap. 23) explicitly introduce the notion of a TEI Abstract Model. The fact is that the notion of an abstract model is used in many formal procedures but this very notion is not formally defined. This ends up in a lot of problems and circularities. We think that we need to have a formalized account of the quasi-formal notion of TEI abstract model, if it has to be of any use other than a sort of regulatory principle.

We do not advocate going back to a monist theory of textuality. Our suggestion to adopt contemporary Semantic Web formalisms to build this abstract conceptual model give us the possibility to have a "foundation" of TEI in a well-defined data model that is not dependent on the notion of a single hierarchical "ordered hierarchy of content object" (OHCO, DeRose et al., 1997), and that can accommodate, at least to some extent, the "pluralities" of textuality.

Structure of the ontology

TEI as a whole is very complex, and its usage is governed by pragmatics and contextual requirements. We acknowledge that it is impossible to reduce to a unique formal semantic definition this fuzzy cloud. Though, we can identify a subset of shared assumptions, a common ground of notions about the meaning of TEI markup and the nature of documents like object: we think that this subset can be the object of an ontological formalization. For various reasons we have adopted the TEI Simple customization (Cummings et al., 2014) as an acceptable approximation of this common ontology. This is not an opportunistic ad hoc choice, as it may seem. TEI Simple in fact has been defined by a group of domain expert that have analyzed the actual usage of TEI markup in some big textual repositories and have selected and organized a set of one hundred or so elements that can describe all the textual features represented by the markup in those documents. This fits perfectly in the definition of a formal ontology development process.

The main design requirements for building our ontology have been the following:

- the ontology must express at the same time an abstract characterization of TEI Simple elements' semantics and an ontological definition of their structural role;
- the ontology must define a precise semantics of the elements having a clear characterization in the official TEI documentation (e.g., the element <p>), while it should relax the semantic constraints if the elements in consideration can be used with different semantic connotations depending on the context (e.g., the element <seg>);
- it must be possible to extend the ontology, reuse it and define alternative characterizations of elements semantics without compromising the consistency of the ontology itself;
- where possible existing ontologies or meta-ontologies must be reused

In accordance with these overall principles we have decided to implement a complex architecture using some pre-existing meta-ontology frameworks to express the meaning of TEI element set by the way of the classes and properties they define. In particular we have adopted:

1) LA-Earmark (Di Iorio, Peroni, Poggi, Vitali, 2011; Peroni, Gangemi, Vitali, 2011), a markup metalanguage, that can express both the syntax and the semantics of markup as OWL assertions, and an ontology of markup that make explicit the implicit assumptions of markup languages. LA-EARMARK is an extension of EARMARK with the Linguistic Act module of the Linguistic Meta-Model that allows one to express and assess facts, constraints and rules about the markup structure as well as about the inherent semantics of the markup elements themselves.

2) Structural Pattern Ontology (Di Iorio, Peroni, Poggi, Vitali, 2014), whose goal is to identify a small number of patterns that are sufficient to express how the structure of digital documents can be segmented into atomic components.

The specification of markup semantics for the various TEI Simple elements is done by means of LA-EARMARK class and properties. The general Earmark class for any markup element is earmark:Element. The <abbr> element is defined as follows:

```
Prefix earmark: <http://www.essepuntato.it/2008/12/earmark#>
Prefix co: <http://purl.org/co/>
Prefix tei: <http://www.tei-c.org/ns/1.0/>
Class: tei:abbr a
 earmark:Element that
 earmark:hasGeneralIdentifier "abbr" and
 earmark:hasNamespace "http://www.tei-c.org/ns/1.0"
```

LA-EARMARK allows us to link particular class of elements with the actual semantics they express. From our point of view there are at least two semantic levels that we explicitly define:

- one concerning the structural behavior of markup that is described by means of the Pattern Ontology (PO);
- the other regarding the intended semantics of an element (e.g., the fact that an element is a paragraph rather than a section, a personal name reference rather than a geographical reference), that is described by TEI Semantics Ontology or by a combination of already existing ontologies.

TEI Semantics Ontology is the core component that gives the actual semantics of TEI elements. Its definition is based on a categorization of the elements of the TEI Simple, based on a refactoring of the TEI model Classes.

The link between the class describing kinds of elements and their related semantic characterization is possible by means of the property "semiotics:expresses". The associations of semantics to markup elements can be contextualized according to a particular agent's point of view, in order to provide provenance data pointing to the entity that was responsible for such specification. This is possible by means of the Linguistic Act Ontology included in LA-EARMARK that allows one to consider all these markup-to-semantics links as proper linguistic acts done by someone.

Conclusions

The work we have done so far is limited to the Simple subset of TEI. We envision some further development:

- Refine the TEI Semantics Ontology component.
- Extend to some other areas of TEI that are suitable for formalization.

We think that in the long term this ontological formalization could become the primary formalization of the TEI encoding schema, independently of any serialization format. Today XML is still the better strategy to encode digital texts in real word projects for many practical reasons. But there is no reason for the TEI to be strictly based on it, as it is *de facto* now. Technical issues should not determine the choice of a formalization language. In the end, we believe that our effort can give a substantial contribution to the TEI to envision the shape of its own future.

Bibliography

Ciotti F. and Tomasi F. (2014). Formal ontologies, Linked Data and TEI semantics. *Journal of the Text Encoding Initiative*, **9**.

Ciotti F., et al. (2015). An ontology for the TEI: one step beyond. *TEI Conference and Members Meeting 2015*. Lyon. http://tei2015.huma-num.fr/en/papers/#acc-16.

DeRose, S. J., Durand, D. G., Mylonas, E. and Renear, A. H. (1997). What is Text, Really?. *Journal of Computer Documentation*, **21**(3): 1–24.

Di Iorio, A., Peroni, S. and Vitali, F. (2009). Towards markup support for full GODDAGs and beyond: the EARMARK approach. *Proceedings of Balisage: The Markup Conference 2009*, Balisage Series on Markup Technologies, Vol. **3**, Montreal, Canada. doi:10.4242/BalisageVol3.Peroni01.

TEI Consortium. (2015). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.9.1. Last updated 15th November 2015. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Cummings, J. et al. (2015). TEI Simple: Power, economy, and a processing model for encoders and developers. *Digital Humanities 2015*. http://dh2015.org/abstracts/xml/CUMMINGS_James_C__TEI_Simple__Power__economy__and/CUMMINGS_James_C__TEI_Simple__Power__economy__and_a_pro.html.

Peroni S., Gangemi A. and Vitali F. (2011). Dealing with Markup Semantics. *I-SEMANTICS 2011 Proceedings*, 111-118. DOI: 10.1145/2063518.2063533.

Di Iorio, A., Peroni, S., Poggi, F. and Vitali, F. (2011). Using semantic web technologies for analysis and validation of structural markup. *Int. J. of Web Engineering and Technology*, **6**(4): 375-98.

Di Iorio, A., Peroni, S., Poggi, F. and Vitali, F. (2014). Dealing with structural patterns of XML documents. *Journal of the American Society for Information Science and Technology*, **65**(9): 1884-900. DOI: 10.1002/asi.23088.

Digital Palaeography: What is digital about it?

Arianna Ciula
arianna.ciula@roehampton.ac.uk
University of Roehampton, United Kingdom

Compared to the tradition of analytical palaeography it builds on, how is digital palaeography transformative? In this paper[1] I will reflect on the emergent meanings and possible research directions of digital palaeography by analysing the last twelve years of approaches and conceptualisations in the field. Moving between a formal and an historically situated analysis, I will show how digital approaches relate to the scholarly tradition of the study of handwriting and writing systems as a whole. Digital palaeography will emerge well positioned to represent the complexity of handwritten objects from an unfamiliar perspective by departing from the structure of the expression of handwriting (text as shape).

Words in context

The adoption and development of digital tools and resources for palaeography and manuscript studies are intertwined with fluctuating cultural attitudes (Busa, 1961; Morando, 1961;[2] McCarty, 2014b; Nyhan et al., 2015). The convergence towards the use of digital[3] coupled with humanities — digital humanities[4] — and therefore palaeography — digital palaeography — denotes the methodology of research being enabled rather than the symbolic form of its objects of analysis or of its outcomes. The scope within which I propose to discuss digital palaeography[5] is hence mainly methodological.

Terms	Semantics	Overall emergent meaning
Computational (e.g. Computational Palaeography)	Process-ability of data	Digital Palaeography ⊃ Computational Palaeography
Digital (e.g. digital vs. analogue)	Representational form of data	
Digital +(e.g. Digital Humanities; Digital Palaeography)	Research methodology	

A formal knowledge of handwriting — e.g. about scripts morphology or terminology used to describe it —[6] Not only palaeography as a discipline — from the 1930s with Bischoff at least, onwards — has subscribed to analytical methods, but more in general the perception of text as divisible entity in opposition to the notion of the ungraspable composition of images has prevailed in humanistic enquiries with few exceptions (e.g. semiotics of art).[7] Computer sciences and image processing techniques offer an *addendum*, a perspective that suits nicely methodological traditions and inclinations of the classificatory minds of palaeographers. Yet, my aim in this paper is to identify any transformative aspects (table 1). So, even if digital palaeography follows a long tradition of analytical approaches to handwriting and an even longer human wish to control writing systems, does it actually affect our conceptualisation of handwriting? How is digital palaeography transformative (in the sense of **digital** +)? or is there a **digital + palaeography**?

Projects rationales and self-narratives

I will summarise some 2004-onwards projects and activities (Ciula, 2004a, 2004b, 2005a, 2005b, 2005c, 2009;DigiPal, 2011-14; *Exploratory Workshops*, 2011)[8] which witness a critical engagement with digital technology, informed by diverse modelling processes and a constructive discussion of the limitations of computational tools. These approaches challenge the notion of palaeography as an auxiliary discipline towards a renewed return to an "integral" (Boyle, 1984) perspective which places palaeography within a wider multi and interdisciplinary framework, linking it with philology, linguistics and even cognitive sciences.[9] The analysis on terminology as well as an overview of practices put the emphasis on a self-reflective approach around the analysis of handwriting beyond strictly computational concerns.

Creating and deflating models

In addition, I will reflect on the potentially productive dilemma digital palaeography approaches suffer from; a dilemma that is made more acute in recent practices compared to the already vivid debates in the 1970s between the historical and "Cartesian" approach (Gumbert, 1976) and in the 90s (Costamagna et al., 1995, 1996). On one hand, palaeographers engaged with the digital are busy building things, what Godfrey-Smith (2009: 108) would call a specific type of models or "imaginary concreta" (creatures in between reality and fiction, between the schematic and the concrete); on the other hand, they are engaged in reflecting about their own practice and in so doing deflate the same models they build.[10]

In a paper questioning the connections between a scriptorium and its products, Ganz states: "searching for the distinctive details of letter forms shared by scribes may risk the application of an over rigid positivism to the study of manuscripts." (Ganz, 2015)

Not to dismiss this warning against positivism in digital humanities,[11] I will claim that what a digital palaeography approach as contextualised earlier on brings to the fore is precisely this awareness and hence the questioning of the 'mechanics' of a topographical or taxonomical analysis. By asking "what is the unit of handwriting? what we considered it to be?," a digital palaeographer is aware that even by getting closer to the supposed materiality of the artefact — e.g. through high resolution images and microscopically segmented image features — she does not lose the lenses palaeographers have being using for interpreting such artefacts in the past, but can consciously decide to put them to test. In this lies one of the paradoxes of the digital and of modelling more in general: it brings perceptual materiality to our scrutiny while taking us away from it. The digital models are used to analyse the objects they are models of, but are also self-reflective tools to question those same models.[12]

Semantics and materiality

Palaeographic research with its focus on the perception of handwriting in morphological terms is a reminder that the handwriting manifests itself as an artefact that is rationalised and divided (hence constructed) only after it is given. By bringing to the fore the picture of writing or the writing as picture (cf. "text as shape" vs. "texts as meaning" in Hassner et al., 2012: 193), palaeographical studies live on the intermedia dimension of handwriting. By this I intend the crossing of the analytical media border between the sensorial and the conceptual qualities of handwriting, form and meaning, visible and invisible, between token and type, *langue* and *parole*. An adapted Hjelmslevian semiotic model of language exemplifies this intermedia dimension of writing.[13]

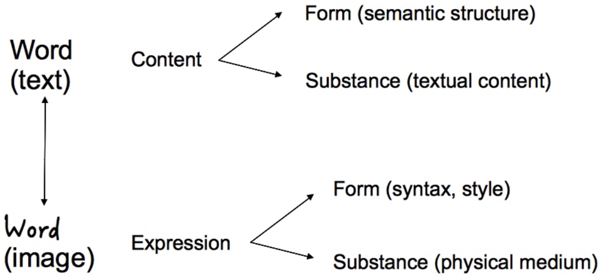

Word (text)
Content
Form (semantic structure)
Substance (textual content)
Word (image)
Expression
Form (syntax, style)
Substance (physical medium)

Figure 1. Interplay between the substance of the expression (the physical medium, the ink on the parchment) and the form of the content (the semantics of the text, its meaning)

These reflections will allow me to sketch the intermedia borders where I think digital palaeography sits. I will build on an unpublished paper (Ciula, 2006) where I used McCloud's (1994) triangular map of visual iconography to represent the relationships between cultural textual objects and their digital (visual) representations both in graphical and in textual forms.[14]

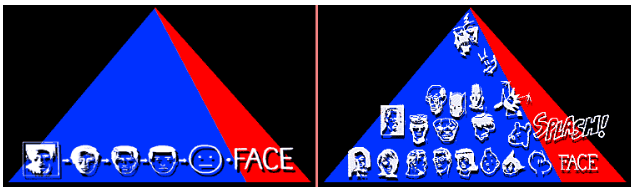


Figure 2: McCloud's Big Triangle or a map of visual iconography (http://archive.is/Vg6OQ)

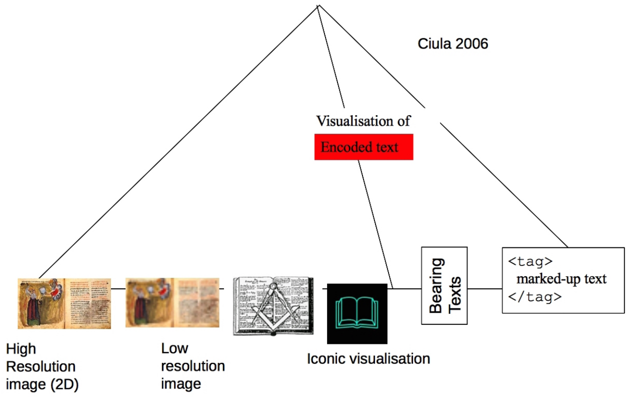


Figure 3. By following the analogy to McCloud's triangle, in Ciula (2006) I showed how the variety of ways of representing a textual object in visual terms can assume both the form of more or less resemblant representations—images of the physical object bearing the handwriting—or their textual counterparts

The visualisation of encoding is an example of structure-oriented visualisation which shows how a graphic rendering of the text does not have to relate unequivocally to features of the textual object as expression (whether form or substance of the expression) but can rather represent one or more supposed structures of the textual content. When made explicit and visible, the structure of the content on the textual-symbolic side can play a fundamental role in the implementation of a thoughtful connection between the image-iconic representation/s of the text and the textual-symbolic content representation/s of a cultural object.[15]

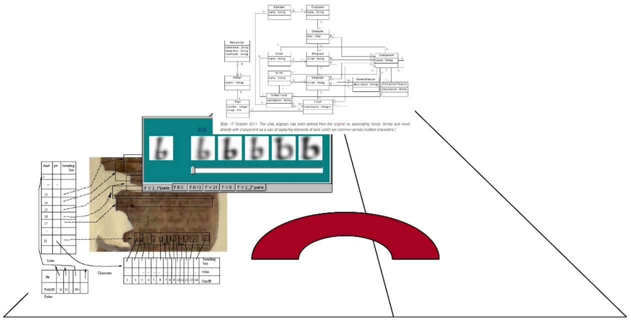


Figure 4. Digital paleographic methods as enhancers of iconic representations of textual artefacts

Following this analysis, digital paleographic methods (inclusive of image processing, image annotation and conceptual models blending morphology and semantics) will be theorised as enhancers of iconic representations of textual artefacts. They can bridge the sensorial/perceptive and structural/conceptual interpretations of handwriting, material and mental knowledge of text, visual and textual, spatial and temporal.

In my 2006 paper, I concluded that for a digital resource to be inspired by and to promote research based on the material/perceptual aspects of a cultural object, a high quality graphical representation of the cultural object is essential but not sufficient. In specular terms to what Buzzetti (2006)[16] says about the symbolic components of textual objects, this paper will argue that the scope of digital palaeography lies in anchoring the structure of the expression of image-texts to the structure of their content, in other words in bridging the "semantic model" of a handwritten source to at least some of the material aspects of the artefact.

Conclusions

Digital palaeography builds on the tradition of analytical palaeography. How is it then transformative or is there a digital + palaeography? Some digital palaeography projects and initiatives, including my own doctoral research, claim to be transformative by advocating for "integral" palaeography and by distancing themselves from a purely computational approach. Some also adopt a self-reflective perspective on the modelling of handwriting in a digital environment, for instance, by testing ontological commitments, categories, classifications of handwriting; in so doing they deflate the models they build. Further, when contextualised within an analysis of the border between form and meaning of handwritten sources, digital palaeography approaches can can be used to connect the structure of expression of handwriting with structures of meaning. A digital model which embeds both structure of expression and structure of content of the handwriting is then theorised as a unique contribution to reconstruct material textuality of cultural artefacts by bridging visual and symbolic elements of texts, spatial and temporal, perception and interpretation. Ultimately, digital palaeography can be transformative by bridging the semantics of written artefacts with their materiality.

Bibliography

Boyle, L. E. (1984). *Medieval Latin Palaeography: A Bibliographical Introduction*. University of Toronto Press.

Busa, R. (1961). L'analisi linguistica nell'evoluzione mondiale dei mezzi d'informazione. *Almanacco Letterario 1962*. Milano: Bompiani, pp. 103–17.

Buzzetti, D. (2006). Biblioteche digitali e oggetti digitali complessi: Esaustività e funzionalità nella conservazione. *Archivi Informatici per Il Patrimonio Culturale*, vol. 114. (Contributi Del Centro Linceo Interdisciplinare «Beniamino Segre»). Roma: Bardi Editore, pp. 41–75 http://web.dfc.unibo.it/buzzetti/dbuzzetti/pubblicazioni/lincei2003.pdf.

Buzzetti, D. and Rehbein, M. (2008). Towards a model for dynamic text editions. In Opas-Hänninen, L. L., Jokelainen, M., Juuso, I. and Seppänen, T. (eds), *Digital Humanities 2008*. Oulu: University of Oulu, pp. 78–81 http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf.

Cecire, N. (2011). When Digital Humanities Was in Vogue. *Journal of Digital Humanities*, **1**(1) http://journalofdigitalhumanities.org/1-1/when-digital-humanities-was-in-vogue-by-natalia-cecire/.

Ciula, A. (2004a). Digital palaeography. *Digital Resources for the Humanities*. Newcastle.

Ciula, A. (2004b). Modelli di scrittura carolina. *Gazette Du Livre Médiéval*, **45**: 27–38.

Ciula, A. (2005a). Digital palaeography: using the digital representation of medieval script to support palaeographic analysis. *Digital Medievalist*, **1**(Spring) http://www.digitalmedievalist.org/journal/1.1/ciula/.

Ciula, A. (2005b). Paleografia e informatica. L'applicazione del software SPI al corpus di manoscritti senesi University of Siena Ph.D.

Ciula, A. (2005c). Un progetto di ricerca. L'applicazione del software SPI ai codici senesi. In Pérez González, C. and Valcárcel Martínez, V. (eds), *Estudios de poesia medieval*. Vitoria: Findación Instituto Castellano y Leonés de la Lengua, pp. 305–22.

Ciula, A. (2006). Cultural Objects in Digital Resources: Imagining the Text. *CLiP 2006: Conference Abstracts: Literatures, Languages and Cultural Heritage in a Digital Worldly*. London: Office for Humanities Communication http://legacy.cch.kcl.ac.uk/clip2006/redist/abstracts_pdfold/paper34.pdf (accessed 28 February 2016).

Ciula, A. (2009). The Palaeographical Method under the Light of a Digital Approach. In Rehbein, M., Schaßan, T. and Sahle, P. (eds), *Kodikologie Und Paläographie Im Digitalen Zeitalter - Codicology and Palaeography in the Digital Age*. (Schriften Des Instituts Für Dokumentologie Und Editorik 2). Norderstedt: Books on Demand (BoD), pp. 219–35 http://kups.ub.uni-koeln.de/2971/.

Ciula, A. and Marras, C. (in press). Circling around texts and language: towards 'pragmatic modelling' in Digital Humanities. *Digital Humanities Quarterly*, **10**(3).

Costamagna, G., Gilissen, L., Gasparri, F. and Pratesi, A. (1995). Commentare Bischoff. *Scrittura e civiltà*, **19**: 321–52.

Costamagna, G., Gilissen, L., Gasparri, F. and Pratesi, A. (1996). Commentare Bischoff. *Scrittura e civiltà*, **20**: 401–07.

Dworkin, C., Morris, S. and Thurston, N. (2012). Information as material http://www.informationasmaterial.org/.

Eyers, T. (2013). The perils of the 'digital humanities': New positivisms and the fate of literary theory. *Postmodern Culture*, **23**(2).

Faulhaber, C. (2015). 28.814 'digital humanities': first occurrence? *Humanist Discussion Group* http://lists.digitalhumanities.org/pipermail/humanist/2015-March/012769.html (accessed 27 February 2016).

Fischer, F., Fritze, C. and Vogeler, G. (eds). (2011). *Kodikologie Und Paläographie Im Digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*. (Schriften Des Instituts Für Dokumentologie Und Editorik 3). Norderstedt: Books on Demand (BoD) http://kups.ub.uni-koeln.de/4337/.

Ganz, D. (2015). Can a Scriptorium Always be Identified by its Products?. In Nievergelt, A., Gamper, R., Bernasconi Reusser, M., Ebersperger, B. and Tremp, E. (eds), *Scriptorium. Wesen - Funktion - Eigenheiten*. (Veröffentlichungen der Kommission für die Herausgabe der mittelalterlichen Bibliothekskataloge Deutschlands und der Schweiz). München: Bayerische Akademie der Wissenschaften, pp. 51–62.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies*, **143**(1): 101–16.

Gumbert, J. P. (1976). A proposal for a Cartesian nomenclature. In Gumbert, J. P. and Haan, M. J. M. de (eds), *Essays Presented to G. I. Lieftinck*, vol. IV. (Miniatures, Scripts, Collections). Amsterdam: A. L. Van Gendt and Co.

Hassner, T., Rehbein, M., Stokes, P. A. and Wolf, L. (2012). *Computation and Palaeography: Potentials and Limits*. Dagstuhl Perspectives Workshop 12382 (Dagstuhl Reports). Dagstuhl: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik http://drops.dagstuhl.de/opus/volltexte/2013/3890/pdf/dagrep_v002_i009_p184_s12382.pdf (accessed 28 February 2016).

Hassner, T., Rehbein, M., Stokes, P. A. and Wolf, L. (2013). *Computation and Palaeography: Potentials and Limits*. Dagstuhl Perspectives Workshop 12382 (Dagstuhl Manifestos). Dagstuhl: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik http://drops.dagstuhl.de/opus/volltexte/2013/4167/ (accessed 28 February 2016).

Hirtle, P. (2000). Editorial. *D-Lib Magazine*, **6**(4) http://www.dlib.org/dlib/april00/04editorial.html.

McCarty, W. (2014a). 27.745 digital knowledge *Humanist Discussion Group* http://lists.digitalhumanities.org/pipermail/humanist/2014-January/011672.html (accessed 28 February 2016).

McCarty, W. (2014b). Getting there from here. Remembering the future of digital humanities Roberto Busa Award lecture 2013. *Literary and Linguistic Computing*, **29**(3): 283–306.

McCloud, S. (1994). *Understanding Comics: The Invisible Art*. First HarperPerennial edition. HarperCollins Publishers.

Montecchi, G. (1998). Gli atlanti dei caratteri tipografici: considerazioni preliminari e propeduetiche dagli scritti di Sigismondo Fanti. In Leonardi, C., Morelli, M. and Santi, F. (eds), *Modi di scrivere: tecnologie e pratiche della scrittura dal manoscritto al CD-ROM. Atti del convegno di studio della Fondazione Ezio Franceschini e della Fondazione IBM Italia. Certosa del Galluzzo, 11-12 ottobre 1996*. (Quaderni di cultura mediolatina 15). Spoleto: Centro italiano di studi sull'Alto Medioevo, pp. 107–30.

Morando, S. (ed). (1961). Le due culture - inchiesta. *Almanacco Letterario 1962*. Milano: Bompiani, pp. 143–44; 314–17.

Nyhan, J., Flinn, A. and Welsh, A. (2015). Oral History and the Hidden Histories project: towards histories of computing in the humanities. *Digital Scholarship in the Humanities*, **30**(1): 71–85.

Pierazzo, E. (2013). A conceptual model of Text, Documents and Work - Part 1 *Elena Pierazzo's Blog* http://epierazzo.blogspot.co.uk/ (accessed 28 February 2016).

Pierazzo, E. (2015). *Digital Scholarly Editing: Theories, Models and Methods*. Aldershot: Ashgate.

Piper, A. (2015). Novel devotions: Conversional reading, computational modeling, and the modern novel. *New Literary History*, **46**(1): 63–98.

Rehbein, M., Schaßan, T. and Sahle, P. (eds). (2009). *Kodikologie Und Paläographie Im Digitalen Zeitalter - Codicology and Palaeography in the Digital Age*. (Schriften Des Instituts Für Dokumentologie Und Editorik 2). Norderstedt: Books on Demand (BoD) http://kups.ub.uni-koeln.de/2939/.

Sahle, P. (5-9 July). What is text? A Pluralistic Approach. *Digital Humanities 2006 Conference Abstracts*. Université Paris-Sorbonne: CATI, pp. 188–90 http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf.

Sahle, P. (2013). Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung. [Preprint-Fassung] Universität zu Köln Ph.D. http://kups.ub.uni-koeln.de/5013/ (accessed 18 February 2016).

Stokes, P. A. (forthcoming). Computing and Palaeography in Theory: Some Historical Context for the Future. In Brookes, S., Rehbein, M. and Stokes, P. A. (eds), *Digital Palaeography*. (Digital Research in the Arts and Humanities). Aldershot: Ashgate.

Stutzmann, D. Écriture médiévale & numérique http://oriflamms.hypotheses.org/ (accessed 28 February 2016).

DigiPal: Digital Resource and Database of Manuscripts, Palaeography and Diplomatic http://www.digipal.eu/ (accessed 27 February 2016a).

Exploratory Workshops: European Science Foundation http://www.esf.org/coordinating-research/exploratory-workshops.html (accessed 27 February 2016b).

Models of Authority: Scottish Charters and the Emergence of Government, 1100–1250 http://www.modelsofauthority.ac.uk/ (accessed 28 February 2016c).

Notes

[1] A preliminary version of this paper was presented at the international workshop *Digital Paleography. Projects, Prospects, Potentialities* (Università Ca' Foscari, Venice, April 10-11, 2014), by invitation of the organisers, Università Ca' Foscari (Venice) and Digital Humanities Lab (DHLAB), EPFL (Lausanne).

[2] My thanks to Willard McCarty for recommending and providing access to these two references.

[3] On the cultural understanding of digital, calling for a wider semantic spectrum that transcends the opposition with analogue, see for example McCarty (2014a). The **plus** sign is a deliberate borrowing of Cecire's reflections on the "problem of the plus" being addictive rather than transformative (McCarty, 2014b: 292): "it should not be possible to have the "plus" without the two terms—"digital" and "humanities"—themselves changing" (Cecire, 2011).

[4] For a recent discussion on the first occurrence of the term digital humanities and its uptake in the first decade of 2000 see Faulhaber (2015).

[5] My thanks to Peter Stokes for having pointed out to me another sense of digital paleography as attributed by Hirtle (2000), who talks about this new "speciality" as the ability to convert obsolete file formats containing digital information (e.g. HTML, JPEG) into current formats. For an overview on the term see Stokes (forthcoming).

[6] Cf. the 15th century treatises on scriptural typologies e.g. as described in Montecchi (1998: 119).

[7] Modern art has exploited this cultural conflict between reading practices that put emphasis on symbolic aspects of written text as opposite to the morphological and sensorial aspects of other artefacts. See for example the works of *Information as Material* (Dworkin, C. et al., 2012) such as the exhibition *Learn to Read Differently* (Northern Gallery of Contemporary Art, Sunderland, August 10– September 23, 2013).

[8] In addition, various events connected to research in digital palaeography took places towards the end of the first decade of 2000 in Europe; in particular, two dedicated symposia on

Codicology and Palaoegraohy in the Digital Age took place in Munich and resulted in two volumes (Rehbein et al., 2009; Fischer et al., 2011). The ESF workshop also had a follow up in the Dagstuhl Perspectives Workshop on *Computation and Palaeography: Potentials and Limits* held in 2012 (see http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=12382) which resulted in the publications Hassner et al. (2012, 2013).

[9] See for instance the ORIFLAMMS project blog (Stutzmann) where the connection to neurosciences is explicit (this project started in 2012).

[10] For a broader adaptation of Godfrey-Smith's "deflationary view" (2009: 115) as a "deflationary account of modelling practices" in digital humanities see Ciula and Marras (in press).

[11] This is a concern beyond digital palaeography with respect, for instance, to trends in distant reading; see for example Eyers (2013).

[12] A similar point with respect to computational modelling of literary novels was made recently by Piper (2015: 68): "We not only gain insights into the specific subset of texts identified by the model, as the model provides the interpretive horizon through which these texts assume new meanings. But we recursively gain insights into the computational model itself through the detailed analysis of the texts it has identified."

[13] While this model can be applied to handwriting and writing alike, the variety of grades of expressions as demarcated in handwriting are more diverse than, for instance, in printed documents.

[14] An alternative circular graph representing the multidimensionality of text encompassing the whole spectrum from physical medium (substance of the expression) to semantics (form of the content) can be found in Sahle's wheel of text (2006, 2013). See also Pierazzo (2013, 2015: 52) who drew two graphs collocating respectively editorial theories and editorial formats on a continuous axis with respect to their relationship with the materiality of text.

[15] Visualisations of dynamic editions (Buzzetti and Rehbein, 2008) integrating textual expression (form or structure of the expression) and semantic model (form or structure of the content) could also exemplify this connection. A concrete example of a project which is attempting at revealing visually the deep connections between the palaeography of texts (handwriting styles of Scottish charters in this case) with the textual content (the representation of authority) is *Models of Authority* (2014-17).

[16] My translation: "The challenge for the representation in digital form of any kind of information and for its adequately exhaustive and functional preservation is therefore given by the possibility of representing the text as a digital complex object and by the ability to reproduce in functional terms the forms of interaction between the structure of the expression and the structure of the content of textual information." (Buzzetti, 2006: 55)

ARLO (Adaptive Recognition with Layered Optimization): a Prototype for High Performance Analysis of Sound Collections in the Humanities

Tanya Clement
tclement@ischool.utexas.edu
University of Texas at Austin, United States of America

Steve McLaughlin
steve.mclaugh@gmail.com
University of Texas at Austin, United States of America

David Tcheng
davidtcheng@gmail.com
University of Illinois Urbana-Champaign

Loretta Auvil
lauvil@illinois.edu
University of Illinois Urbana-Champaign

Tony Borries
tony.borries@gmail.com
University of Illinois Urbana-Champaign

Introduction

Beyond simple annotation and visualization tools or expensive proprietary software, open access software for accessing and analyzing audio is not widely available for general use by the humanities community. Speech recognition algorithms in projects such as MALACH (Multilingual Access to Large spoken ArCHives) are often not built as Web-accessible interfaces for broader audiences. Analysis and visualization software such as PRAAT, which is used by linguists, and Sonic Visualizer, which is often used by music scholars, are desktop tools that typically allow users to focus on one file at a time, making project-sharing difficult for collaborative research and classroom projects. In bioacoustics, researchers use Raven (from the Cornell Lab of Ornithology) and Avisoft (expensive, proprietary software), which perform well with clean data from a single animal. Most of these tools are either not used in multiple domains or with large collections and none of them do well with noise or with multiple signals. As a result of these factors, humanists have few opportunities to use advanced technologies for analyzing large, messy sound archives. In response to this lack, the School of Information (iSchool) at the University of Texas at Austin (UT) and the Illinois Informatics Institute (I3) at the University of Illinois at Urbana-Champaign (UIUC) are collaborating on the HiPSTAS (High Performance Sound Technologies for Access and Scholarship) project. A primary goal of

HiPSTAS is to develop a research environment that uses machine learning and visualization to automate processes for describing unprocessed spoken word collections of keen interest to humanists.

This paper describes how we have developed, as a result of HiPSTAS, a machine learning system called ARLO (Adaptive Recognition with Layered Optimization) to help deal with the information challenges that scholars encounter in their attempt to do research with unprocessed audio collections.

ARLO (Adaptive Recognition with Layered Optimization) Software

ARLO was developed with UIUC seed funding for avian ecologist David Enstrom (2008) to begin exploring the use of machine learning for data analysis in the fields of animal behavior and ecology. ARLO software was chosen as the software we would develop through HiPSTAS primarily because it extracts basic prosodic features such as pitch, rhythm, and timbre that humanities scholars have called significant for performing analysis with sound collections (Bernstein, 2011; Sherwood, 2006; Tsur, 1992).

Filter Bank Signal Processing and Spectrogram Generation

ARLO analyzes audio by extracting features based on time and frequency information in the form of a spectrogram. The spectrogram is computed using band-pass filters linked with energy detectors. The filter bank approach is similar to using an array of tuning forks, each positioned at a separate frequency, an approach that is thought to best mimic the processes of the human ear (Salthouse and Sarpeshkar). With filter banks, users can optimize the trade-off between time and frequency resolutions in the spectrograms (Rossing, 2001) by choosing a frequency range and 'damping factor' (or damping ratio), a parameter that determines how long the tuning forks 'ring.' By selecting these features, users can optimize their searches for a given sound. For these reasons,

Machine-Learning Examples and the ARLO API (Application Programming Interface)

In ARLO, examples for machine learning are audio events that the user has identified and labeled. Audio events comprise a start and end time such as a two-second clip, as well as an optional minimum and maximum frequency band to isolate the region of interest. Users label the examples of interest (e.g., "applause" or "barking"). Other control parameters such as damping factor are also provided for creating spectrograph data according to optimal resolutions for a given problem. The algorithm described below retrieves the features of the tag according to the user's chosen spectra and framing size (e.g., two frames per second, each 0.5 seconds) from the audio file through the ARLO API.

ARLO Machine-Learning Algorithms: IBL (Instance-Based Learning)

The ARLO IBL algorithm finds matches by taking each known classified example and "sliding" it across new audio files looking for good matches based on a distance metric. The average of the weighted training set classes determines prediction probability. The number of match positions considered per second is adjustable and is set to the spectral sample rate. In addition to simple spectra matching, a user can isolate pitch and volume traces, compute correlations on them, and weight the different feature types when computing the overall match strength. This allows the user to weight spectral information that might correspond to such aspects as pitch or rhythm. In the IBL algorithm, accuracy is measured using a simulation of the leave-one-out cross-validation prediction process described above.

Use Case: Finding Applause in PennSound Poetry Performances

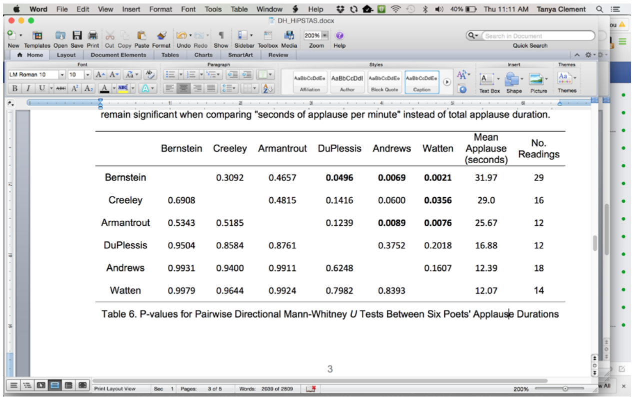
Humanities scholars have identified the sound of applause as a significant signpost for finding patterns of interest in recorded poetry performances. Applause can serve as a delimiter between performances, indicating how a file can be segmented and indexed. Applause can also serve as a delimiter between the introduction to a performance and the moment when a performance has ended and a question-and-answer period has begun, both of which indicate contextual information such as the presence of people who might not appear in traditional metadata fields (Clement and McLaughlin, 2015). A means for quantifying the presence of applause can also lead researchers to consider more in-depth studies concerning the relationship between audience responses and a poet's performance of the same poem at different venues as well as the differing responses of audiences at the same venue over the course of a poet's career or perhaps as a point of comparison between poets. Examples of these results are described below.

For this use case, we ingested approximately 30,257 files remaining (5374.89 hours) from PennSound into ARLO. We chose 2,000 files at random, manually examined them for instances of applause, and chose one instance of applause per recording until we had an example training set of 852 three-second tags, including 582 3-second instances of non-applause (3492 0.5-second examples) and 270 3-second instances of applause (1620 0.5-second examples). Optimization for the IBL test went through 100 iterations. As a result of this optimization process, we used the following parameters for both tests: 0.5-second spectral resolution; 0.5 damping factor; 0.8 weighting power (for IBL); 600 Hz minimum frequency; 5000 Hz maximum frequency; 64 (IBL) and 256 (Weka) spectral bands; spectral sampling rate of 2 (i.e., half-second resolution).

Preliminary Results

We first evaluated our models using cross-validation on the training data. Using the leave-one-out approach, the IBL classifier achieved an overall accuracy of 94.52% with a 0.5 cut-off classification threshold. After comparing 676 configurations, we found that the optimal approach was using IBL with Hann smoothing over 14 windows (7 seconds). The accuracy for this configuration was 99.41%.

In our initial analysis of classification data, we identified significant differences between measured applause durations for six poets, each with more than ten readings in the evaluation set. Table 6 presents the results of pairwise single-tailed Mann-Whitney (Mann and Whitney, 1947) *U* tests of applause durations that have been predicted by our IBL classifier. The alternative hypothesis states that the performer in the left column tends to receive more applause than the corresponding one listed in the top row. Results that are significant at the p<0.05 level appear in bold, with the counts and overall means of each set of observations provided in the right two columns. It appears, for instance, that the poet Rae Armantrout tends to receive more applause than either Bruce Andrews or Barrett Watten. These two differences remain significant when comparing "seconds of applause per minute" instead of total applause duration.



remain significant when comparing "seconds of applause per minute" instead of total applause duration.

	Bernstein	Creeley	Armantrout	DuPlessis	Andrews	Watten	Mean Applause (seconds)	No. Readings
Bernstein		0.3092	0.4657	**0.0496**	**0.0069**	**0.0021**	31.97	29
Creeley	0.6908		0.4815	0.1416	0.0600	**0.0356**	29.0	16
Armantrout	0.5343	0.5185		0.1239	**0.0089**	**0.0076**	25.67	12
DuPlessis	0.9504	0.8584	0.8761		0.3752	0.2018	16.88	12
Andrews	0.9931	0.9400	0.9911	0.6248		0.1607	12.39	18
Watten	0.9979	0.9644	0.9924	0.7982	0.8393		12.07	14

Table 6. P-values for Pairwise Directional Mann-Whitney *U* Tests Between Six Poets' Applause Durations

Table 1. P-values for Pairwise Directional Mann-Whitney *U* Tests Between Six Poets' Applause Durations

Discussion and Future Work

This is preliminary work in an ongoing attempt to create a virtual research environment for analyzing large collections of audio. These data warrant further scrutiny, however, since multiple factors might be skewing the results. First, recording technologies have changed over time and as a result some earlier recordings likely include more noise and thus more false positives. Second, editing practices and event formats can vary widely between venues and over time. Finally, recordings that are included in the PennSound archive represent curation decisions that may favor certain kinds of performers over others. Part of the challenge is to determine what use such analysis might serve for scholars in the humanities and in other fields. Some of the HiPSTAS participants have written about their experiences using ARLO in their research (MacArthur, 2015; Mustazza, 2015; Sherwood, 2015; Rettberg, 2015), but we are interested in feedback from multiple user communities including linguists and scientists who have recorded too much sound data for traditional forms of analysis and processing (Servick, 2014). Furthermore, the IBL algorithm produced promising results, but could be improved with more training data, for example. Or, in extended experiments in which users wish to increase the accuracy of the model, we could develop a voting mechanism on the predictions by comparing the models. Users could validate newly identified examples and include them as new training examples, building each model again on the new data. We are currently working on further testing the models and developing a means for these iterative approaches.

Bibliography

Bernstein, C. (2011). *Attack of the Difficult Poems: Essays and Inventions*. Chicago: University Of Chicago Press.

Bioacoustics Research Program. (2014). Raven Pro: Interactive Sound Analysis Software (Version 1.5). Ithaca, NY: The Cornell Lab of Ornithology.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, **5**(9/10): 341-45.

Cannam, C., Landone, C. and Sandler, M. (2010). Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. In *Proceedings of the ACM Multimedia 2010 International Conference*.

Council on Library and Information Resources and the Library of Congress. (2010). *The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age*. Washington DC: National Recording Preservation Board of the Library of Congress.

Enstrom, D. A. and Ward, M. P. (2008). Sex specific song repertoires in Northern Cardinals: mutual assessment and the occurrence of female song. *The 12 International Behavioral Ecology Meetings*, Ithaca, NY.

Greene, M. A. and Meissner, D. (2005). More Product, Less Process: Revamping Traditional Archival Processing. *The American Archivist*, **68**(2): 208–63.

Hall, M., et al. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, **11**(1).

MacArthur, M. (2015). Monotony, the Churches of Poetry Reading, and Sound Studies. *PMLA*. Forthcoming.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, **18**(1): 50–60.

Mustazza, C. (2015) The noise is the content: Toward computationally determining the provenance of poetry recordings using ARLO. *Jacket2*. Retrieved from https://jacket2.org/

commentary/noise-content-toward-computationallydetermining-provenance-poetry-recordings

Nelson-Strauss, B., Gevinson, A. and Brylawski, S. (2012). The Library of Congress National Recording Preservation Plan. Washington, DC: Library of Congress.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, **2**(1): 37–63.

Rettberg, E. (2015). Hearing the Audience. *Jacket2*, **26**. Retrieved from http://jacket2.org/commentary/hearing-audience.

R Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved fromhttp://www.R-project.org

Rossing, T. D., and Moore, F. R. (2001). *The Science of Sound* **(3rd edition.)**, San Francisco: Addison-Wesley.

Salthouse, C. D. and Sarpeshkar, R. (2003). A Practical Micropower Programmable Bandpass Filter for Use in Bionic Ears. *IEEE Journal Of Solid-State Circuits*, **38**(1): 63-70.

Servick, K. (2014). Eavesdropping on Ecosystems. *Science Magazine*. doi: 343.6173: 834–837.

Sherwood, K. (2006). Elaborate Versionings: Characteristics of Emergent Performance in Three Print/Oral/ Aural Poets. In *Oral Tradition*, **21**(1): 119-47.

Sherwood, K. (2015). Distanced sounding: ARLO as a tool for the analysis and visualization of versioning phenomena within poetry audio. *Jacket2*. Retrieved from https://jacket2.org/commentary/distanced-sounding-arlo-toolanalysis-and-visualization-versioning-phenomena-within-poetr.

Smith, A., Allen, D. R. and Allen, K. (2004). Survey of the State of Audio Collections in Academic Libraries. Washington, DC: Council on Library and Information Resources.

Tsur, R. (1992). *What Makes Sound Patterns Expressive?: The Poetic Mode of Speech Perception*. Duke University Press.

Contextualized Integration of Digital Humanities Research: Using the NeMO Ontology of Digital Humanities Methods

Panos Constantopoulos
p.constantopoulos@dcu.gr
Athens University Of Economics and Business; Digital Curation Unit, Athena Research Center

Lorna M. Hughes
Lorna.Hughes@glasgow.ac.uk
University of Glasgow

Costis Dallas
c.dallas@dcu.gr
Panteion University; University of Toronto; Digital Curation Unit, Athena Research Center

Vayianos Pertsas
vpertsas@aueb.gr
Athens University Of Economics and Business; Digital Curation Unit, Athena Research Center

Leonidas Papachristopoulos
l.papachristopoulos@dcu.gr
Digital Curation Unit, Athena Research Center

Timoleon Christodoulou
christodoulout@aueb.gr
Athens University Of Economics and Business

The advent of digital infrastructures for arts and humanities research calls for deeper understanding of how humanists work with digital resources, tools and services as they engage with different aspects of research activity: from capturing, encoding, and publishing scholarly data to analyzing, visualizing, interpreting and communicating data and research argumentation to co-workers and readers. Digitally enabled scholarly work, and the integration of digital content, tools and methods, present not only commonalities but also differences across disciplines, methodological traditions, and communities of researchers. A significant challenge in providing integrated access to disparate digital humanities (DH) resources and, more broadly, in supporting digitally-enabled humanities research, lies in empirically capturing the context of use of digital content, methods and tools. This paper presents recent and ongoing work on the development of NeMO, an ontology of digital methods in the humanities, and its deployment for the development of a knowledge base on scholarly work.

Several attempts have been made to develop a conceptual framework for DH in practice. In 2008, a project funded by the UK's Arts and Humanities Research Council, the AHRC ICT Methods Network, based at King's College, London, developed a taxonomy of digital methods in the arts and humanities. This was the basis for the classification of over 200 digital humanities projects funded by the AHRC in the online resource arts-humanities.net. This taxonomy was subsequently modified by Oxford University as the basis for the classification of digital humanities initiatives at the University (Digital Humanities at Oxford). Other initiatives to build a taxonomy of Digital Humanities include TaDiRAH and DH Commons. From 2011 to 2015 the European Science Foundation funded the Network for Digital Humanities in the Arts and Humanities (NeDiMAH). This Network was established to develop a better understanding of the practice of DH across Europe, and ran over 40 activities structured around key methodological areas in the humanities (digital representations of space and time; visualisation; linked data; creating and using large scale corpora; and creating editions). Through these activities, NeDiMAH gathered a snapshot of the practice of digital humanities in Europe, and the impact of digital methods on research. A key output of NeDiMAH is NeMO: *the NeDiMAH Ontology of Digital Methods in the Arts and Humanities*. This ontology of digital methods in the humanities has been built as a framework for understanding not just the use of digital methods, but also their relationship to digital content and tools. The development of an ontology, rather than a taxonomy, stands in recognition of the complexity of the digital humanities landscape, the interdisciplinarity of the field, and the dependencies that impact the use of digital methods in research.

NeMO was developed by the Digital Curation Unit (DCU), IMIS-Athena Research Centre, in collaboration with NeDiMAH, as a conceptual framework capable of representing scholarly work in the humanities, addressing aspects of intentionality and capturing the diverse associations between research actors and their goals, activities undertaken, methods employed, resources and tools used, and outputs produced, with the aim of obtaining semantically rich structured representations of scholarly work. It is grounded on earlier empirical research through semi-structured interviews with scholars from across Europe, which focused on analysing their research practices and capturing the resulting information requirements for research infrastructures. Its intellectual foundations lie in earlier work of the DCU on conceptualizing and modelling scholarly activity in the arts and humanities, conducted within the Preparing DARIAH,DYAS / DARIAH-GR, and EHRI projects, and manifested in the Scholarly Research Activity Model (SRAM), an ontological representation of scholarly information activity drawing from cultural-historical activity theory and process modelling, and compatible with CIDOC's Conceptual Reference Model (CIDOC CRM, ISO 21127:2006).

Architecturally, NeMO adopts a three layer structure, spanning from abstract/general to concrete/special concepts, to provide a flexible framework suitable to the multidisciplinarity of DH. Its top tier concepts (*Actor, Activity, Object*) provide a general reasoning frame, and function as semantic links to reference ontologies such as CIDOC CRM. These abstract notions are specialized in the second layer by way of domain-specific concepts covering every aspect of scholarly work: *Methods* employed in activities of various degrees of complexity or taught in *Courses*, *Tools* used, *Information Resources* taken as input or produced as output, *Groups/Organizations* or *Persons* participating in various roles, *Goals* addressed, *Topics* covered, etc. Furthermore, in this second layer, several semantic relations capturing the context of the aforementioned core concepts allow for modeling scholarly work through four complementary perspectives: (1) Process-related, centered around the concept of *Activity* and capturing temporal and spatial aspects; (2) Methodological, centered around the *Method* concept and capturing "how" aspects; (3) Agency-related, centered around the *Actor* and *Goal* concepts and capturing "who" and "why" aspects; and (4) Resource-related, centered around the *Information Resource* concept and covering "what" aspects of scholarly work.

In the third layer of NeMO, fine-grained notions supporting domain-specific detailed descriptions are represented as specializations of second layer concepts. Respective vocabularies are organized as SKOS thesauri. More specifically, controlled vocabularies of lexical terms are structured hierarchically under the concepts of *ActivityType, MediaType, InformationResourceType, TopicKeyword, ActorRole, SchoolOfThought* and *Discipline*, which are specializations of the *Type* concept of the second layer, and are used for characterization/classification in parallel to ontological classification. The role of these taxonomies is, thus, twofold: (1) as a vocabulary of terms that can be used for flexible tagging of the objects of interest; (2) as entry points for the alignment, or mapping, of terms from NeMO to terms from other existing taxonomies. The latter enables integration with related work, as well as effective use of these taxonomies as documentation instruments or entry points for content in NeMO knowledge bases. For instance, the *ActivityType* taxonomy is organized in five hierarchies roughly corresponding to Unsworth's "cholarly primitives", and offers a flexible tagging system for modelling the intentionality of actors, scope adherence of activities, or purpose of use of tools and methods. On the other hand, mappings through broader/narrower term relations from the *ActivityType* terms to terms of other method taxonomies, including TaDiRAH, Oxford ICT and DH Commons, allow using those taxonomies transparently within NeMO.

The development of NeMO contributes to the work of the Digital Methods and Practices Observatory Working Group of DARIAH (DiMPO), as well as of Europeana Research within the Europeana Cloud project, providing an intellectual foundation for the analysis of evidence on arts and humanities scholarly activities and needs with regard to digital resource access across Europe. The relevance of the ontology to the DH community was validated through interviews and web surveys, to elicit information needs and patterns in working practices among humanities researchers, as well as two workshops in which these patterns were explored through use cases contributed by researchers. The evidence collected demonstrates that NeMO addresses adequately the knowledge representation needs manifested there. A variety of complex associative queries articulated by researchers in these workshops were also collected, demonstrating the potential of NeMO as an effective mechanism for information extraction and reasoning with regard to the use of digital resources in scholarly work; queries were encoded in SPARQL, a language appropriate for exploiting the serialization of NeMO in RDF Schema (RDFS), thus highlighting the benefits of its potential use as a knowledge base schema.

A prototype implementation of the above functionalities provides an easy to use demonstration of NeMO's potential. Users can articulate queries in structured English, without prior knowledge of any specific query language, using an intuitive user interface offering dynamic feedback of suggestions based on the conceptual schema. Input to the knowledge base is also supported by the same mechanism, guiding the user according to relations and classes provided by the model. A use case will be presented by way of example.

In sum, NeMO offers a well-founded conceptualization of scholarly work, which can function as schema for a knowledge base containing information on scholarly research activity, including goals, actors, methods, tools and resources involved. NeMO can thus be useful to researchers by (a) helping them find information on earlier work relevant for *their own* research; (b) supporting goal-oriented organization of research work; (c) facilitating the discovery of yet uncharted paths with regard to resources, tools and methods suitable for particular contexts; and, (d) promoting networking among researchers with common interests. Additional benefits for research groups include support for better project planning by explicitly representing links between goals, actors, activities, methods, resources and tools, as well as assistance for discovering methodological trends, future directions and promising research ideas. Furthermore, for funding organisations, research councils, etc., NeMO can (a) provide a bird's eye view of funded scholarly activities; (b) enable the systematic documentation of research projects; (c) support evaluation of proposals, monitoring and control of project work, and validation of project outcomes.

Planned improvements include the development of mechanisms for providing recommendations based on semantically related instances and for the semi-automatic population of the knowledge base, as well as specialization of core classes and addition of new terms in Type taxonomies to reflect developments in DH scholarship.

Bibliography

Benardou, A., Constantopoulos, P. and Dallas, C. (2013). An approach to analyzing working practices of research communities in the humanities, *International Journal of Humanities and Arts Computing*, **7**: 105-27 (Edinburgh University Press).

Benardou, A., Constantopoulos, P., Dallas, C. and Gavrilis D. (2010). Understanding the information requirements of arts and humanities scholarship: implications for digital curation. *International Journal of Digital Curation*, **5**(1).

Hughes, L. M., Constantopoulos, P. and Dallas, C. (2014). Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines, *The New Companion to Digital Humanities*, (Eds.) Schreibman, S. and Siemens, R. (Oxford, Blackwell).

Hughes, L. M., and Ell, P. (2013). Digital Collections as Research Infrastructure, *From Evolution to Transformation: Research Infrastructures and Scholarly Research*. Special issue of the *International Journal of Humanities and Arts Computing*, (Edinburgh University Press).

Hughes, L. M. (2011). ICT Methods for digital collections research *Digital Collections: Use, Value and Impact*. London: Facet.

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. *Humanities Computing, Formal Methods, Experimental Practice Symposium. 5100.*

Frequency and Meaning

Hugh Craig
hugh.craig@newcastle.edu.au
University of Newcastle, Australia

1. Introduction

Computational stylistics now has thirty years or so of publications and has been celebrated as one of the success stories of Digital Humanities (McCarty, 2014: 289). It brings together statistical methods and literary analysis, inferring meaning from the frequency of literary features. In this paper I explore this connection between frequency and meaning, and consider some of the objections which have been made to the statistical approach to style.

Much research in linguistics focuses on individual sentence-level structures. Stylistics introduces a new dimension of extension and cumulation, placing a net of continuing co-occurrence over a language sequence. The added dimension of time, or extent, opens up the analysis of meaning not only in the instance itself but in the series that it forms with other instances of the same feature. Computational stylistics takes a routinely quantitative approach to this cumulative aspect.

Critics have laid down significant challenges to this frequentist approach. They have questioned whether language features are really countable, whether frequency matters in meaning, whether the inevitable choice of features to count undermines the objectivity of the results, and whether quantitative results can ever usefully relate the text to any wider context.

2. Are language features countable?

The first key enabling assumption of computational stylistics is that the language features being counted are homogenous. In his 1970s articles attacking stylistics, Stanley Fish argues that meaning is constructed by the reader at the moment of reading and concludes that stylistics is therefore an invalid practice. There is no meaning in the word on the page, so it is pointless to count instances of the word, of combinations of words, or of any other language feature (Fish, 1973; Fish and Graham, 1979). John Frow, likewise, argues that in literary study features are not stable or commensurate but relational, so counting them is pointless (cited in Bennett, 2009: 287).

3. Is frequency meaningful?

Many would also question the relation between frequency and salience. It seems unwise to assume that an unusual accumulation of a feature is necessarily noticed by writer or reader. Fish in a recent online post critiquing digital humanities argues that only patterns intended by the author are worth discussing (Fish, 2012). Different frameworks influence the noticeability of language elements, and a single instance may be highly salient, and a cluster of instances may pass without any conscious reaction.

4. Function words

Considering function words as a basis for counting helps counter these objections. Computational stylistics has a natural alliance with function words. Function words lend themselves to computation since they are easy for a machine to recognise and appear regularly and in large numbers, offering opportunities for analysis by statistical methods whose power is well established in other domains. On the other hand, computation has a special benefit for function words analysis because counting on a scale not possible for the unaided reader makes it possible to reveal hitherto latent patterns in the behaviour of these words.

Function words do not have a semantics in the usual sense: *if* has a structural function rather than a meaning. The stylistic import of the word only becomes clear in repetition. By contrast, lexical words are rich in meaning in the individual instance and do not necessarily achieve any cumulative effect through a series. Function words bear traces of larger structures and hence, though not salient in themselves, their frequencies bear meaning as indexes to wider discourse orientations. They help show how a language feature can be sufficiently homogenous to justify counting, and how frequencies can have a literary dimension.

There are two other important objections to consider: the possible bias arising from the fact that a judgement has to be made about which features to count, and the difficulty of relating patterns found within a corpus to extra-textual factors.

5. Features have to be chosen, so results are arbitrary

Tony Bennett points out that researchers have to choose the units to count in – there are no "given units" – and argues that this choice has a necessary influence on results, which undermines any claims to objectivity (2009: 290, 291).

This is a fundamental critique of quantitative study, i.e. of any quantitative study. The logical extension would be that the choice of units always determines the results, so there can be no surprises and nothing new can be learned. It is easy to show that there are cases where this is not so. If we ask, do women write differently from men? - we have a way of validating the units: if the pattern of use of a given unit shows a significant difference in a balanced and commensurate sets of samples of the writing of women and the writing of men, then it does not matter how the unit was chosen. Here we have an external basis, the difference between two objectively based classes, on which to discard some units and accept others. Then there are cases of classification, e.g. by author and by date. We can seek markers of the classes, check them with known members of the classes, and then apply them to disputed cases. We have an objective way of validating the units, so we don't care much about where they came from.

6. Formalism

Computational stylistics begins with textual features, focuses on finding patterns in their use, provides striking visualisations of the patterns, and then struggles to relate the patterns to extra-literary events. The textual data is well defined, easy to explore, and with the help of statistics it can be shown that there are robust structures within it. The world of possible causation beyond is hard to limit and hard to quantify. If there is (say) a consistent and marked increase in the Shannon Entropy of the language of Victorian novels from early in the period to late, how could that be described in terms of the reading experience? And how could that be related to the forces acting on the novel? Computational stylistics is lop-sided: very well developed on the textual side, but weak - tentative and fragmentary - in relating statistical findings to the extra-textual world. Another way of saying this is to call computational stylistics formalist. In this sort of approach the evidential force of the explanation for a pattern will always be less than that for the pattern itself. However, it is only fair to point out that in this it is in the same situation as other literary methods. A literary effect may be demonstrable, but its genesis in composition, and the larger forces to which it relates, are always matters of judgement and selective contextualisation. The text is available, even if dauntingly complex, but the conditions which made it possible have to be painstakingly and always speculatively recreated. It is easier to show that Hamlet changes in the course of his play than that this observed change relates to Early Modern beliefs about the typical course of melancholia.

7. Conclusion

Computational stylistics has proved itself in the realm of classification. In this area the methods can be thoroughly tested and success or otherwise can be demonstrated. There are some well-established and significant findings, leading to a reassessment of some commonplaces such as the downplaying of authorship as a factor in style (Egan, 2014). This presents a problem for those who think that counting literary features is inherently unsafe, that frequencies in language cannot have any real force, and that all feature choice is fatally arbitrary. Beyond classification, though, these objections still have some force, and a new one intrudes, the argument that computational stylistics is disablingly formalist. Computational stylistics now needs to produce findings in more properly stylistic areas of the same weight as its justly celebrated classification ones, findings which match the style within a corpus to the world beyond it. Only then will we be confident that frequency in literary language is linked to meaning, and that computational stylistics has the methods to do justice to this link.

Bibliography

Bennett, T. (2009). Counting and Seeing the Social Action of Literary Form: Franco Moretti and the Sociology of Literature. *Cultural Sociology*, **3**: 277-97.

Egan, G. (2014). What Is Not Collaborative About Early Modern Drama in Performance and Print? *Shakespeare Survey*, **67**: 18-28.

Fish, S. E. (1973). What Is Stylistics and Why Are They Saying Such Terrible Things About It?. In Chatman, S. (ed) *Approaches to Poetics: Selected Papers from the English Institute*. New York: Columbia University Press, pp. 109-52.

Fish, S. E. (2012). *Mind Your P's and B's: The Digital Humanities and Interpretation*. New York Times, Opinionator. http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/ (accessed 6 March 2016)

Fish, S. E. and Graham, J. F. (1979). *What Is Stylistics and Why Are They Saying Such Terrible Things About It? Part II. Boundary*, **28**: 129-45.

McCarty, W. (2014). Getting There from Here. Remembering the Future of Digital Humanities: Roberto Busa Award Lecture 2013. *Literary and Linguistic Computing*, **29**: 283-306.

Creating An EpiDoc Corpus for Ancient Sicily

James Cummings
James.Cummings@it.ox.ac.uk
University of Oxford, United Kingdom

Jonathan Prag
jonathan.prag@merton.ox.ac.uk
University of Oxford, United Kingdom

James Chartrand
jc.chartrand@gmail.com
Open Sky Solutions

Introduction

This paper introduces the TEI P5 XML – EpiDoc corpus of inscriptions on stone for ancient Sicily, I.Sicily. The project is one of the first attempts to generate a substantial regional corpus in EpiDoc. The project is confronting a number of challenges that may be of wider interest to the digital epigraphy community, including those of unique identifiers, linked data, museum collections, mapping, and data conversion and integration, and these will be briefly outlined in the paper which will concentrate on the conversion and technical development of the project.

Technical Background of I.Sicily

I.Sicily is an online, open access, digital corpus of the inscriptions on stone from ancient Sicily.[1] The corpus aims to include all texts inscribed on stone, in any language, between approximately the seventh century BC and the seventh century AD. The corpus currently contains records for over 2,500 texts, and when complete is likely to contain c. 4,000. The corpus is built upon a conversion from a legacy dataset of metadata in MS Access to EpiDoc TEI XML. The XML records are held in an eXist database for xQuery access, and additionally indexed for full-text search using SOLR/Lucene. The corpus and related information (museum list, bibliography) are published as Linked Data, and are manipulated through a RESTful API. The records are queried and viewed through a web interface built with AngularJS and jQuery javascript components. Mapping is provided in the browser by the Google Maps API, and ZPR (Zoom, Pan, Rotate) image- viewing is provided by the IIIP image server.

At the time of writing, the main conversion routine is being refined, and the epigraphic texts are being collated for incorporation into the records. An ancillary database of museum collections in Sicily has been constructed and bibliography is held in a Zotero library. Extensive search facilities will be provided, including map-based and bibliographic searching. Individual inscriptions and individual museums will both be provided with URIs, as will personal names and individuals; places will be referenced using Pleiades, epigraphic types, materials, and supports using the EAGLE vocabularies.

The motivations for I.Sicily

The existing epigraphic landscape in Sicily is extremely diverse in two primary regards: on the one hand, the island has a very mixed cultural and linguistic make-up, meaning that the epigraphic material is itself extremely varied, with extensive use throughout antiquity of both Greek and Latin, as well as Oscan, Punic, Sikel, and Hebrew[2]; on the other hand, the publication of this material has a very uneven record and despite an excellent pre-twentieth century tradition, the existing corpora are far from complete and the ability of key journals such as SEG or AE to keep pace with local publication has been limited. A limited number of museum-based corpora have been published in recent decades (for Catania, Palermo, Messina, and Termini Imerese, as well as the material from Lipari), but this has not greatly improved the overall situation. The combination of these two factors already means that locating, identifying, or working with a Sicilian inscription, or its publication record, is extremely challenging for anyone without extensive experience of the material. I.Sicily has been conceived in the hope of improving the situation in all these areas.

Multilingualism

Sicily is traditionally described as a 'melting pot', the 'crossroads of the Mediterranean'. The situation created by basic technologies such as Unicode and TEI P5 EpiDoc XML mean that there is now no reason not to be language agnostic in the inclusion of material. The opportunities and possibilities offered by these technologies are considerable, since, for example, searching can be made language specific or language neutral. One obvious area where Sicilian studies are currently hampered by this disciplinary partitioning is in the study of onomastics. The Lexicon of Greek Personal Names records most instances of Greek names for the island, but Sicily is no less rich in non-Greek names (Latin and others), and at present there is no onomasticon for the island.[3] Simply by the marking-up and indexing of all names in the island's inscriptions, I.Sicily will have generated a powerful tool for future study.

Identification and Bibliography

The PHI database of Greek inscriptions has a rich record of Greek texts, but again is text only and limited in outputs.[4] SEG references are available for 733 inscriptions on stone and AE references for 328 (data taken from the I.Sicily database and based upon comprehensive manual trawls of SEG and AE). One major aim of I.Sicily, therefore, is to generate unique identifiers for each inscription - the I.Sicily number, in the form ISic 1234 maintained as URIs, of the form: http://sicily.classics.ox.ac.uk/isicily/inscriptions/1234. I.Sicily is well placed to do this since its initial dataset is primarily a bibliographic concordance of the lapidary inscriptions of Sicily. One of the associated outputs of the project will therefore be an online bibliography for Sicilian epigraphy, and an online Zotero library has already been created with over 700 records which are referenced in the EpiDoc. A locally cached version of the bibliography will be presented at the I.Sicily site to facilitate detailed bibliographic searching (including the identification of inscriptions by publication) and to allow the generation of customised concordances.

Location, location, location

I.Sicily is actively generating rich geo-data for the individual inscriptions, both for the original findspot/provenance and the current location (whether museum-based, on-site, or elsewhere), and we aim to provide map-based searching for inscriptions, as well as text-based searching by ancient and modern place-names. In addition to full listing wherever possible of both ancient and modern place names for epigraphic provenance, we are working to provide detailed location information for each find-spot and current location, through a combination of library and map-based research and the use of autopsy and GIS recording. At present geo-data is being recorded in two forms, both through the use of explicit geographical locations in the form of longitude and latitude records in decimal degree form, and through the use of Pleiades URI references wherever possible.[5] We are committed to the long-term use of Pleiades as our primary reference for ancient places, and to that end we aim to update and improve the Pleiades data for Sicilian locations, in particular name data and sub-locations, in conjunction with the editing of the I.Sicily records.

Translations

The creation and availability of translations is a major goal of the EAGLE project and its collaborators, and I.Sicily is no less committed to that ambition.[6] Translations are very rarely available for any of the published Sicilian inscriptions. It is obvious that the inclusion of translations will make the material much more accessible to a wider audience both of students and the general public. Equally, provision of translations will add to the value of the database as a resource for museums and others curating the inscriptions recorded in the database. To that end, a long-term ambition of I.Sicily is to include translations wherever possible in both English and Italian. We see this as one obvious area where public contribution ('crowd-sourcing') will be invaluable.

Limitations and future ambitions

The scale of the enterprise, and the available resources, mean that in its current form the project has limited itself to inscriptions engraved on stone (the coverage of rupestral inscriptions/graffiti and of inscriptions painted on stone/plaster is regrettably uneven). However, there is no reason in principle not to extend coverage in future to include inscriptions on other materials. Similarly, although the current project does not include a programme to mark up linguistic features of the texts, the commitment to the long-term maintenance of the corpus and the open availability of the underlying XML records means that such a project would be entirely possible in the future.

It is our long-term ambition that I.Sicily might become the default location for the publication and dissemination of Sicilian inscriptions; in the shorter term, we hope that it will serve as valuable portal in the world of Sicilian epigraphy and of ancient world open linked data, greatly improving the accessibility of Sicilian epigraphy and so enriching the study of the 'crossroads of the Mediterranean'.

Bibliography

Gulletta, M. I. (ed.) (1999), *Sicilia Epigraphica*. Atti del convegno internazionale, Erice, 15-18 ottobre 1998, 2 vols.

Orlandi, S., Santucci, R., Casarosa, V. and Liuzzo, P. M. (eds.) (2014). Information Technologies for Epigraphy and Cultural Heritage, *Proceedings of the First EAGLE International Conference, Rome 2014*, Sapienza Università Editrice. Published online at: http://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf (accessed 5 March 2016)

Prag, J. R. W. (2002). Epigraphy by numbers: Latin and the epigraphic culture in Sicily. In Cooley, A. E. (ed.), Becoming Roman, Writing Latin?: Literacy And Epigraphy In The Roman West (Journal of Roman Archaeology Supplementary Series), *Journal of Roman Archaeology*, pp. 15-31.

Tribulato, O. (ed.) (2012). *Language and linguistic contact in ancient Sicily*, Cambridge: Cambridge University Press.

Notes

[1] The corpus will be mounted at http://sicily.classics.ox.ac.uk/isicily/ by the time of DH2016, but is currently on a development server.

[2] Recent overview of much of the linguistic tradition in Tribulato 2012; and of the epigraphic material in Gulletta 1999.

[3] See http://www.lgpn.ox.ac.uk/

[4] See http://noapplet.epigraphy.packhum.org/regions/1156

[5] See http://pleiades.stoa.org/

[6] See Orlandi et al. 2014: Part II.

Sustainable publishing – Standardization possibilities for Digital Scholarly Edition technology

Alexander Czmiel
czmiel@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

After decades of building digital resources for humanities research, such as Digital Scholarly Editions (DSE), and making them available to researchers and the broader public, we are at the point where many of these resources can be connected to one another and are more and more accepted by the scholarly community. However, we also experience the challenge to maintain all the various Digital Scholarly Editions which were built on a diverse base of different technologies. This is especially complex as Digital Scholarly Editions are "living" objects. On the one hand that means that the content can be extended and refined continuously. Hence they are never finished. On the other hand the technological basis must be kept accessible, secure and running. Those two processes can be summarized under the term "data curation".

If we assume that a Digital Scholarly Edition not only consists of the marked up texts, mostly XML documents, but also of another layer on top of the XML documents, the functionality layer – all the interactive parts, the visualizations and the different views on the texts, indexes or other research material, such as images or audio documents – it is obvious that data curation can become an unlimited complex task. This functionality layer provides an enormous additional benefit to the texts. A Digital Scholarly Edition can be seen as a tool which is used to analyze the XML documents, thus as part in the research process which must be preserved to reproduce research results which often cannot be achieved without the functionality layer.

A Digital Humanities resource usually undergoes a typical life cycle and is built by a team of team members with a variety of competences that are needed for each task:

1. Analysis of the sources to be edited (humanities scholars)

2. Requirement Engineering (the whole project team)

3. Design of the data or document model, choosing what standards to use (scholars, database-, markup-, metadata-specialists)

4. Choosing, adopting, and/or developing software tools for transcription, editing and publishing (software developers, scholars)

5. Installing and maintaining development servers and web servers (system administrators)

6. Conceptual design and implementation of the web publication of the Digital Scholarly Edition (web designer, web developer, scholars)

7. Preparation for long term access and archiving (documentation- and metadata-specialists)

8. Service support and maintenance after project finished (data curators)

At each step of this life cycle decisions are made which have impact on the subsequent steps. The first two steps of the list constitute the foundation on which the whole Digital Scholarly Edition is built on, from the data model over the choice of software tools until the publication as well as data curation.

Digital Scholarly Editions are sufficiently described from a methodically point of view regarding the document and text modeling (Pierazzo 2015, Sahle 2013). An analytical description from the technological point of view still is a desideratum. To make a comprehensive data curation possible a technological publishing concept which uses standardized components is needed. Such a concept can consist of standards for a formal project documentation, a description of the used technologies, the provided interfaces and APIs, a design paradigm for typical user interaction tasks, and many more. Standards on the data- and metadata-layer are broadly accepted and in use – one example are the Guidelines of the Text Encoding Initiative (TEI – http://www.tei-c.org) – but they are still missing for the functionality layer.

A high standard critical Digital Scholarly Edition can only be built in a sustainable way and be maintained when it follows technological standards which still have to be developed. The paper will present a first tiny step of a proposal for a minimal standard from the technological point of view of a Digital Scholarly Edition. It focuses on experiences made during the last ten years working on XML-based Digital Scholarly Editions built with certain tools, such as eXistdb (http://exist-db.org). Hence the proposed solution cannot be valid for all the different kinds of Digital Humanities scholarly resources.

A possible next step towards such a formal description could be to package those XML-documents together with the source code of the functionality layer in a standardized self-descriptive format. An option for this task could be the EXPath Packaging System (http://expath.org/modules/pkg/), which works well for XML-based Digital Scholarly Editions and is widely used by Digital Scholarly Editions which are published using eXistdb. The main purpose of such a packaging system is not connectivity or interoperability rather than maintenance and data curation. The packaging system can be extended gradually to a technological publishing format which incorporates the aforementioned aspects such as a project description format.

A possible formal project description format for the documentation will consist of the following information:

- The name of the project and all involved institutions and persons.
- The status of the project: planned, work in progress, published, or finished.
- The applied technologies and standards.
- The licenses, which are used for research data, source code and other components such as fonts, audio or video documents.
- Information about where to find the source code, if the source code is available under an open source license.
- Information about provided APIs and other interfaces to retrieve the research data and metadata in various formats (XML, JSON etc.) or get structured information about persons or places to be processed further in other contexts (In case of a correspondence edition metadata about the letters should be prepared in the Correspondence Metadata Interchange Format (CMIF) to be reused by http://correspsearch.bbaw.de).
- Contextual details about the data producers, how data are collected etc. (More at Faniel 2015)
- Canonical citation rules and instructions for persistent referencing of current parts and older versions of the research data.
- A standardized change log, which can be evaluated by other services.

Of course this list can be just a first suggestion and does not provide all the information that can be given about a project. The project description must be accessible under a standardized URL (e.g. http://home.of.project/api/projectdescription) and can be serialized in different formats, such as XML or JSON, for further processing. That would allow a Digital Scholarly Edition to be registered at a central directory where all information and updates of various Digital Scholarly Editions which follow the same publishing model are collected automatically. Such a central directory does not exist yet. Currently existing directories collect information manually and describe projects externally, so changes and updates are harder to track.

The success of such a publishing model depends on pragmatic usage possibilities and a critical mass of Digital Humanities scholars and projects who publish their Digital Scholarly Edition using this publishing model. It is difficult to find a standardized, generic approach in the world of Digital Scholarly Editions as every project encounters a different set of problems and a different set of uses. Thus it is important as developers to not make too many assumptions about the nature of a project and further the development of a technological publishing standard in continuous exchange with the scholarly community and in very small steps which take into account the diversity across the Humanities.

Bibliography

Faniel, I. (2015). *Data Management and Curation in 21st Century Archives*, http://hangingtogether.org/?p=5375> (accessed 6th March 2016).

Pierazzo, E. (2015). *Digital Scholarly Editing, Theories, Models and Methods.* Ashgate.

Sahle, P. (2013). *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, Bände, Norderstedt: Books on Demand 2013, **3**.

Visual Patterns Discovery in Large Databases of Paintings

Isabella di Lenardo
isabella.dilenardo@epfl.ch
DHLAB - EPFL, Switzerland

Benoit Seguin
benoit.seguin@epfl.ch
DHLAB - EPFL, Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
DHLAB - EPFL, Switzerland

The digitization of large databases of works of arts photographs opens new avenue for research in Art History. For instance, collecting and analyzing painting representations beyond the relatively small number of commonly accessible works was previously extremely challenging. In the coming years, researchers are likely to have an easier access not only to representations of paintings from museums archives but also from private collections, fine arts auction houses, art historian photo collections, etc. However, the access to large online databases is in itself not sufficient. There is a need for efficient search engines, capable of searching painting representations not only on the basis of textual metadata but also directly through visual queries. In this paper we explore how convolutional neural network descriptors can be used in combination with algebraic queries to express powerful search queries in the context of Art History research.

Context and Method

This project is part of project called *Replica*, conducted in collaboration with the Cini Foundation in Venice. This project is based on two parallel developments, the digitization of the Cini Foundation's photo library, a collection of about a million photographs of paintings, engravings, plastic arts and architecture (1300 - 1900) and the creation of a dedicated search engine allowing for searches for visual patterns in this database. As the digitization is currently ongoing, the results reported in this paper are performed on a subset of only 39 000 paintings. However, the progressive densification of the database, including a large set of so-called minor painters, should, in the coming months, unfold the full discovery potential of this search engine.

The field of visual pattern recognition has been recently transformed by the surprising performances of so-called deep learning approaches using Convolutional Neural Networks (CNN). CNN are multi-layers architectures used for supervised learning, especially for object classification. Each layer is representing an operation on the previous layer: convolution layer, fully-connected layer, pooling layer, regularization layer, etc. These networks have many parameters (filter parameters, fully-connected weights) that can be learned via backpropagation in a supervised manner. Traditionally, the input (first) layer is a full raster image and the output (last) layer is a vector representing the score of the input image for each class. By showing the network some labeled images and comparing the network's output to the desired label, one can update the parameters of the model.

The theory of deep neural methods have been known for decades, and were already successfully applied with the first convolutional neural networks in the 90s to digit recognition (LeCun et al., 1989). However, their computational complexity and the necessity of important amount of training data have seen them being ignored for a long time. With very large datasets available like *ImageNet*, and GPU computation being more accessible, there has been a sudden surge of interest in deep methods (Deng et al., 2009).

In 2012, a convolutional neural network shattered the competition in a difficult 1000 class object recognition challenge, attaining the impressive result of a top–5 error of 15.3% compared to 26.3% for the runner-up (Krizhevsky et al., 2012). Ultimately, this work had an important impact on the machine vision community starting the so-called deep learning revolution. A clear manifestation of this trend was that just a year later at the next iteration of this object recognition challenge, almost every entry was based on CNNs as well.

Despite being trained to recognize a precise set of classes. It has been observed that some of the learned parameters of the CNNs will most likely be similar across different datasets. For instance, the first convolutional layer usually learns various edge detectors and basic filters. Some researchers have evaluated the representative power of CNN trained for a specific task to other problems. Using a model that outperformed the others in 2012 on the *ImageNet* data, results seemed extremely promising, suggesting that task transfer is possible (Donahe et al., 2014).

To calculate the descriptors of our search engine, we use a pre-trained Convolutional Neural Networks similar to the one described in (Donahe et al., 2014). Each painting of our database is associated with 1000 features, corresponding to the last convolutional layer of the pre-trained Convolutional Neural Network. These features are thought to represent high-level characteristics directly usable for the classification tasks. Through this process each painting is associated with a single point in a high dimensional space. When a single image query is sent to the search engine, the results are simply shown, ranked by their distance to the query.

However, similarity between paintings could not be the results of single homogenous distance. To enable the users to specify the kind of similarity they want to explore, a more refined language has been introduced. Searches take the form of *algebraic formulas* in which the user can add or subtract examples. For performing such searches, we use a binary support vector machine (kernel Radial Basis Function). In the cases where no negative examples are provided, a one-class-support vector machine is used (in the case of a query with a single image, this corresponds to a simple nearest neighbor algorithm). The rest of the paper shows examples of such algebraic queries and the corresponding results.

Examples of queries and results

The classic principles for classifying visual similarities in Art History include various dimensions like recurrence of particular pictorial patterns or common compositional structures. As a query illustration, the first criteria of classification chosen is the search for common 'dominant and multiple pictorial motif' in the composition. One classic example in this typology is the *Still life*, featuring for instance only a large bouquet of flowers. The development of this subject has a long history, from the late sixteenth century, before arriving at its codification during the seventeenth-century. The results of a query with Juan de Arellano's *Still life of flowers* (Fig. 1) include other famous interpreters of the genre, almost identical in composition and also close in terms of chronological and pictorial influences: de la Corte, Snyders, Casteels. However, there are also seventeenth-century painters, Gentileschi, Régnier, Bonito, Vouet, who, while not painting Still lifes, are characterized by the same tonality of *chiaroscuro* typical of this precise moment in history of art. Without further information the similarities found by a single image query include various families of resemblances combining pictorial patterns and color tones.

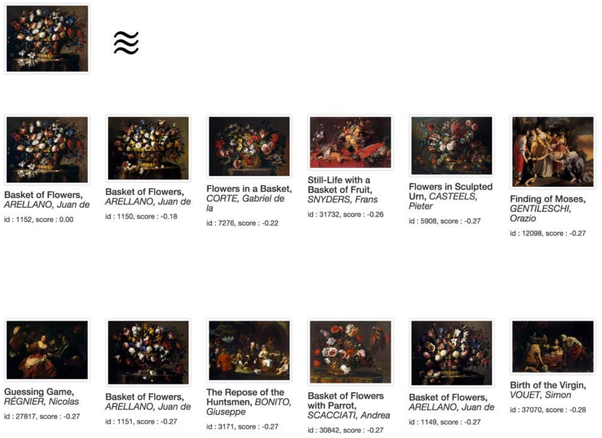

≈
Basket of Flowers, *ARELLANO, Juan de*
id : 1152, score : 0.00
Basket of Flowers, *ARELLANO, Juan de*
id : 1150, score : -0.18
Flowers in a Basket, *CORTE, Gabriel de la*
id : 7276, score : -0.22
Still-Life with a Basket of Fruit, *SNYDERS, Frans*
id : 31732, score : -0.26
Flowers in Sculpted Urn, *CASTEELS, Pieter*
id : 5908, score : -0.27
Finding of Moses, *GENTILESCHI, Orazio*
id : 12098, score : -0.27
Guessing Game, *RÉGNIER, Nicolas*
id : 27817, score : -0.27
Basket of Flowers, *ARELLANO, Juan de*
id : 1151, score : -0.27
The Repose of the Huntsmen, *BONITO, Giuseppe*
id : 3171, score : -0.27
Basket of Flowers with Parrot, *SCACCIATI, Andrea*
id : 30842, score : -0.27
Basket of Flowers, *ARELLANO, Juan de*
id : 1149, score : -0.27
Birth of the Virgin, *VOUET, Simon*
id : 37070, score : -0.28

Figure 1 : A query of a Still Life of Flowers by Juan de Arellano returns several paintings with flowers but also other subjects

To focus only on the pictorial motif of flowers excluding any paintings with figures, we subtract one of the paintings by Gentileschi to the initial query (Fig. 2). We obtain all the 'key painters' in this genre including for instance Daniel Seghers. He does not paint a real Still life but flowers around a sacred figure, the Virgin, one of the first subjects, probably invented by Jan Brueghel the Elder. It is probably from this initial subject that evolves the *Still Lifes with flowers*. So, in this case, the algebraic query recovered the evolution of a specific pictorial motif with its significant variations during the Seventeenth century.

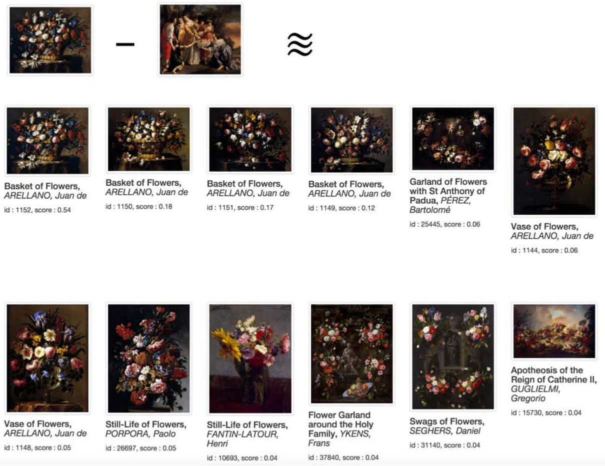

–
≈
Basket of Flowers, *ARELLANO, Juan de*
id : 1152, score : 0.54
Basket of Flowers, *ARELLANO, Juan de*
id : 1150, score : 0.18
Basket of Flowers, *ARELLANO, Juan de*
id : 1151, score : 0.17
Basket of Flowers, *ARELLANO, Juan de*
id : 1149, score : 0.12
Garland of Flowers with St Anthony of Padua, *PÉREZ, Bartolomé*
id : 25445, score : 0.06
Vase of Flowers, *ARELLANO, Juan de*
id : 1144, score : 0.06
Vase of Flowers, *ARELLANO, Juan de*
id : 1148, score : 0.05
Still-Life of Flowers, *PORPORA, Paolo*
id : 26697, score : 0.05
Still-Life of Flowers, *FANTIN-LATOUR, Henri*
id : 10693, score : 0.04
Flower Garland around the Holy Family, *YKENS, Frans*
id : 37840, score : 0.04
Swags of Flowers, *SEGHERS, Daniel*
id : 31140, score : 0.04
Apotheosis of the Reign of Catherine II, *GUGLIELMI, Gregorio*
id : 15730, score : 0.04

Figure 2 : By subtracting to the flower painting the Finding of Moses by Orazio Gentileschi, only painting featuring flower are returned

Another criteria of classification in Art History are structural analogies between compositions (Gombrich, 1960). Structure of the composition is understood here in a geometrical sense, with the reoccurrence of similar geometrical patterns in various paintings. The formal analysis of paintings have classically focused on such kinds of structural similarity (Focillon 1964; Didi-Huberman 1996).

The *The Gallery of Archduke Leopold* painted by David Teniers the younger, in 1639, is also known to have four different variants. This painting is considered a reference of a long tradition of paintings subjects featuring cabinet of art lovers and collectors, a well studied genre (Findlen 1996; van der Veen 1993). A single query returns variations of the same painting (Staatsgalerie Schleissheim, Münich; Prado, Madrid; Kunsthistorisches Museum, Wien), but also painting featuring squares within squares (Fig. 3). For instance on *The Ambassadors depart by Vittore Carpaccio and Baptism of St Libertus by Colijn de Coter,* squares are on the floor or on the wall. To refine further our search and find the "good neighborliness" (Warnke 2000; Freedberg 1989) of David Teniers structure, we can try to exclude these two examples by subtracting them to the initial query. Results of such a query exclude now interior scenes featuring geometrical squares but now include various scenes of the *Passion of the Christ* organized as sequences of "squares", non featuring any elements in perspective (Fig. 4). In a third attempt, we can now exclude those by substracting them to the initial query and reinforce the focus on search by adding variant of the first painting by Teniers. Indeed, all the first results feature now paintings with galleries of collectors with examples of the most important authors of the genre, thus facilitating the study of their mutual influence (Fig. 5).

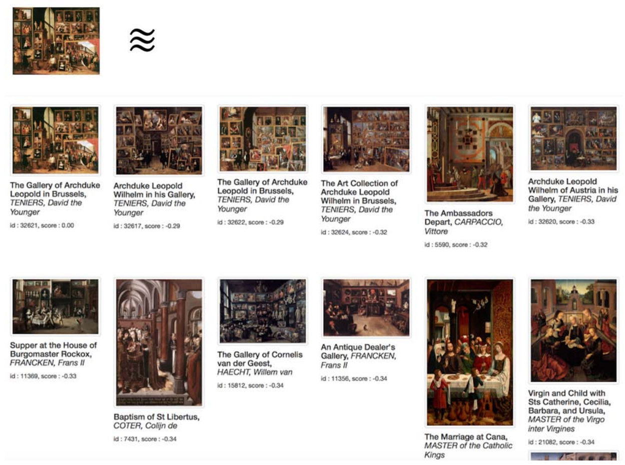

≈
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32621, score : 0.00
Archduke Leopold Wilhelm in his Gallery, TENIERS, David the Younger
id : 32617, score : -0.29
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32622, score : -0.29
The Art Collection of Archduke Leopold Wilhelm in Brussels, TENIERS, David the Younger
id : 32624, score : -0.32
The Ambassadors Depart, CARPACCIO, Vittore
id : 5590, score : -0.32
Archduke Leopold Wilhelm of Austria in his Gallery, TENIERS, David the Younger
id : 32620, score : -0.33
Supper at the House of Burgomaster Rockox, FRANCKEN, Frans II
id : 11369, score : -0.33
Baptism of St Libertus, COTER, Colijn de
id : 7431, score : -0.34
The Gallery of Cornelis van der Geest, HAECHT, Willem van
id : 15812, score : -0.34
An Antique Dealer's Gallery, FRANCKEN, Frans II
id : 11356, score : -0.34
The Marriage at Cana, MASTER of the Catholic Kings
Virgin and Child with Sts Catherine, Cecilia, Barbara, and Ursula, MASTER of the Virgo inter Virgines
id : 21082, score : -0.34

Figure 3: A query with the The Gallery of Archduke Leopold by David Teniers the younger (1639) gives a first results four variants of the same painting by the same author. The following results include various kind of painting which same some similarities with the initial query but are not representing the same subject

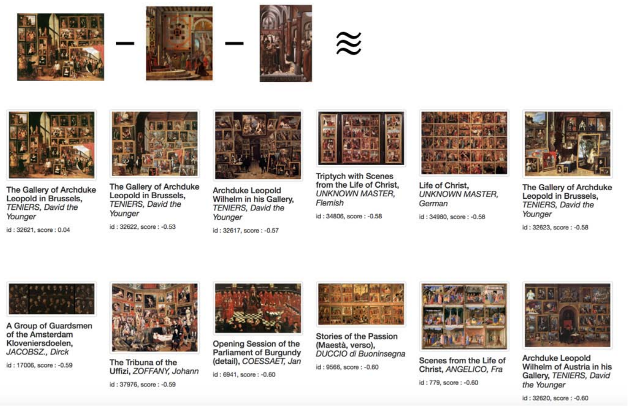

–
–
≈
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32621, score : 0.04
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32622, score : -0.53
Archduke Leopold Wilhelm in his Gallery, TENIERS, David the Younger
id : 32617, score : -0.57
Triptych with Scenes from the Life of Christ, UNKNOWN MASTER, Flemish
id : 34806, score : -0.58
Life of Christ, UNKNOWN MASTER, German
id : 34980, score : -0.58
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32623, score : -0.58
A Group of Guardsmen of the Amsterdam Kloveniersdoelen, JACOBSZ., Dirck
id : 17006, score : -0.59
The Tribuna of the Uffizi, ZOFFANY, Johann
id : 37976, score : -0.59
Opening Session of the Parliament of Burgundy (detail), COESSAET, Jan
id : 6941, score : -0.60
Stories of the Passion (Maestà, verso), DUCCIO di Buoninsegna
id : 9566, score : -0.60
Scenes from the Life of Christ, ANGELICO, Fra
id : 779, score : -0.60
Archduke Leopold Wilhelm of Austria in his Gallery, TENIERS, David the Younger
id : 32620, score : -0.60

Figure 4: When The Gallery of Archduke Leopold is subtracted with The Ambassadors depart by Vittore Carpaccio and Baptism of St Libertus by Colijn de Coter a series of paintings only containing hierarchy of embedded squares are returned. The formula has isolated a specific characteristic in the feature space when the presence of a multiple squares is the most specific trait

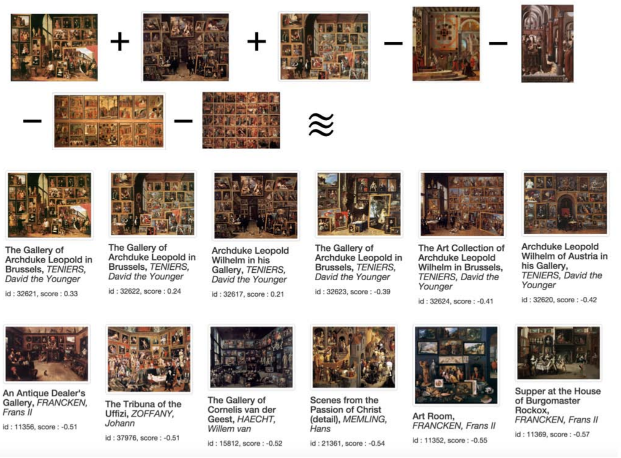

+
+
–
–
–
–
≈
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32621, score : 0.33
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32622, score : 0.24
Archduke Leopold Wilhelm in his Gallery, TENIERS, David the Younger
id : 32617, score : 0.21
The Gallery of Archduke Leopold in Brussels, TENIERS, David the Younger
id : 32623, score : -0.39
The Art Collection of Archduke Leopold Wilhelm in Brussels, TENIERS, David the Younger
id : 32624, score : -0.41
Archduke Leopold Wilhelm of Austria in his Gallery, TENIERS, David the Younger
id : 32620, score : -0.42
An Antique Dealer's Gallery, FRANCKEN, Frans II
id : 11356, score : -0.51
The Tribuna of the Uffizi, ZOFFANY, Johann
id : 37976, score : -0.51
The Gallery of Cornelis van der Geest, HAECHT, Willem van
id : 15812, score : -0.52
Scenes from the Passion of Christ (detail), MEMLING, Hans
id : 21361, score : -0.54
Art Room, FRANCKEN, Frans II
id : 11352, score : -0.55
Supper at the House of Burgomaster Rockox, FRANCKEN, Frans II
id : 11369, score : -0.57

Figure 5: In order to search specifically for the paintings containing paintings, the two painting representing the stories of the Passion of the Christ can be subtracted. Most results now feature paintings in which paintings are present

Perspectives

Pattern recognition methods have made extremely impressive progresses in the recent years, thanks to the advances of convolutional neural network and the advent of very large databases of images. As Art History deals with the study of the migration of patterns, this is surely a great opportunity for designing new tools to search through large databases of paintings photographs. This paper is a first examination of the typologies of search use cases that could be envisioned combining convolutional neural network features and simple algebraic formulas. This initial study illustrates that this new way of expressing queries allows for the incremental definition of various kinds of "similarities" between paintings. Convolutional neural networks features manage to capture many dimensions of similarity between paintings, including composition, colors and also common iconographic elements. Combined with a simple language for expressing the specificity of the traits that the user looks for, it could enable new powerful search tools that may in turn have important impact on history of art studies.

Bibliography

Deng, J., Dong, W., Socher, R., Li, L.-J. Li, K. and Fei-Fei, L. (2009). ImageNet: *A large-scale hierarchical image database, IEEE Conf. Comput. Vis.* Pattern Recognit.

Didi-Huberman, G. (1996). Pour une antropologie des singularités formelles. Remarque sur l'invention warburgienne. *Genèses*, **24**: 145-63.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014). DeCAF: *A Deep Convolutional Activation Feature for Generic Visual Recognition,* Int. Conf. Mach. Learn, pp. 647–55.

Findlen, P. (1996). *Possessing Nature - Museums, Collecting &*

Scientific Culture in Early Modern Italy. Reprint. Berkeley: University of California Press.

Focillon, H. (1964). *Vie des formes: suivie de l'"Eloge de la main"*, PUF, Paris.

Freedberg, D. (1989). *The Power of Images: Studies in the History and Theory of Response*, Chicago-London.

Gombrich, E. H. (1960). *Art and Illusion: A Study in the Psychology of Pictorial Representation*, Pantheon Books, New York.

Krizhevsky, A. Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Computation*, pp. 541–51.

Van der Veen, J. (1993). *Galerij en Kabinet, vorst en burger. Schilderijen Collectie in de Nederlanden, In* Bergvelt, E., Meijers, D. J. and Rijnders, M. eds. *Verzamelen van Rariteitenkabinet tot Kunstmuseum,* Heerlen, pp. 145-64.

Warnke, M. (2000)*Aby Waburg - Der Bilderatlas. Mnemosyne*, Akademie, Berlin.

Urban Youth and Community Media: A Digital Place-Making Process in Vanuatu

Thomas Dick
tom.dick@scu.edu.au
Further Arts, Vanuatu

Sarah Doyle
info@furtherarts.org
Further Arts, Vanuatu

In this paper, we focus on media (and multimedia) production as a spatial practice of ni-Vanuatu youth. We explore the contexts and practices of youth media production in a contemporary postcolonial urban society and how organizational forces shape these practices. Articulation theory is used as a framework for thinking about the way that young urban ni-Vanuatu are negotiating community, institutional, and professional obligations as well as leveraging opportunities for producing media that demonstrate new modes of relating to place. The authors draw on long term coactivity and co-performance (Conquergood and Johnson, 93) via participant engagement with artists and producers from Vanuatu. Both authors have lived and worked in Vanuatu for over five years and much of our engagement with the young producers is through our roles with Further Arts – an NGO based in Port Vila. We employ a radical empiricism that blurs the boundaries between observer and observed – a form of dialogic performance that embraces and complicates diversity, difference, and pluralism (Conquergood and Johnson, 93). Madison describes it as living in "embodied engagement of radical empiricism, to honor the aural/oral sounds that incorporate rather than gaze over" (168, emphasis in original). A range of visual media productions is analyzed, principally video productions. We also draw on interviews conducted with members of Nesar Studio – a community access media production studio located at Further Arts – who constitute an emergent social category in Vanuatu, that is: young independent media producers or the 'youth media crew'.

The contemporary inflections of the precolonial ni-Vanuatu relationship to place reveal contingent openings for indigenous people to transform their world through ontologically liberating participation in media production (and consumption). From another perspective, multinational companies in the resource extraction (mining, forestry, fishing), agriculture, tourism, and creative industries, operate in ways that diminish Oceanic ontologies and subjectivities through processes that undermine local agency, knowledge, wisdom, value systems and biocultural diversity. But a dualistic framing such as this reinforces tired false binaries. It sets us a dangerous path to navigate through essentialism, exceptionalism and reductionism that, even if successfully negotiated ultimately leads to intellectual dead ends. There needs to be another way of understanding the: dynamic trajectories of visiting and returning, assembling and reassembling; the pluralism of Oceanian actuality, beyond the static divisions of rural villages and urban towns. In this paper we extend Clifford's reading of Hall's articulation theory (Hall; in Grossberg "On Postmodernism and Articulation: An Interview with Stuart Hall"), and demonstrate ways that "the partial entanglements of indigenous and local societies in global structures are not simply the world system's unfinished business. They have their own roots and trajectories." (Clifford, 475). We explore these roots and trajectories through – "a praxis of spatial articulation" (Tawa, 49). Dick has documented elsewhere forms of ni-Vanuatu cultural production and expression that reflect a chorographic engagement with place (Tawa; Olwig; Maxwell; Dick).

In this paper, we extend these ideas exploring the performative inflections of media production in a praxis of youth participation and spatial articulation. We will make visible the roots and trajectories of ni-Vanuatu youth and "the forces (the articulations) that create and maintain identities that have real concrete effects" (Slack, 126). Recognizing the joining and the un-joining, the assembling and the reassembling, the creating and the recreating, the articulating and re-articulating, imbues this approach with its efficacy (Slack; Clifford) and interdisciplinary approaches can converge to refine the way that we understand the contemporary media world (Horst, Hjorth and Tacchi).

The paper begins with a description of the major historical and cultural forces and tensions that influence the locative identities of people in urban Vanuatu, integrated with a contextualization – a re-articulation – of the Vanuatu mediascape from the perspective of young urban ni-Vanuatu producers. We explore the specific case study of the establishment of Nesar Studio and how this studio is facilitating the emergence of young independent (and interdependent) producers, or the 'youth media crew' (henceforth YMC), as differentiated from the category of media consumers. The dissolution of observer and observed in the project, that is, our embedded-ness in the data, creates an opportunity for a deeply integrative analysis of contemporary indigenous engagements with media and place and multilateral strategies of articulation and de-articulation. Working in this radically empiricist way, it is often difficult to balance the requirements of the academy and the expectations of the community; it is difficult to "shape the data", as it were, when one is both in and part of the data (Grossberg We Gotta Get out of This Place: Popular Conservatism and Postmodern Culture 55-56). Thus we have integrated discussion and analysis into the presentation of data to which it directly relates.

Despite major challenges, the young producers at Nesar Studio assert the importance, and indeed their ownership, of the structure and its organizational relationships – in particular with FA. This is evidenced in the extent to which they worked to maintain their momentum after massive damage from Cyclone Pam. 'Nesar' means 'nasara' in a local language – nasara is the Bislama term used throughout Vanuatu to talk about the ceremonial meeting place of a village for the intergenerational transmission of kastom knowledge and wisdom (through song, dance, art and other practices) is transmitted. Taking on this word and its connotations, Nesar Studio becomes a digital urban nasara in an age of increased use and access to telecommunications and media platforms as ways to transmit messages and knowledge. Providing the community with education on these tools is a powerful means to enact change through engaging people with their rights.

Developing the capacity of young media professionals in Vanuatu is not a simple process. Stakeholders must navigate divergent interests and compete for resources within an environment where the digital media and creative industries are poorly understood by state mechanisms and supported in an ad-hoc fashion by development partners at both national and regional levels. The bodies that do exist to advocate for and stimulate this sector, including the Vanuatu Cultural Centre, Secretariat of the Pacific Community, and the Pacific Arts Association (amongst others) have not been overly influential or consistently effective. This means that the role of shaping and nurturing the capacity of young media professionals rests in the hands of civil society actors, either in cooperation with, or struggling alongside, the dominant media and communications companies.

Functions, processes, and forces historically familiar to ni-Vanuatu communities are deployed in the articulation of youth (id)entities through media production practices in urban settings. Only a small number of the more talented and adept media crew members of Nesar Studio have been able to penetrate the industry in an independent and professional capacity, acquiring employment in other cultural and media agencies or as contractors for national and international organizations. While these opportunities are few and far between, they require a certain level of application and assertiveness on the part of the media producer, something that Nesar Studio aspires to for its members, but perhaps does not emphasise enough since its media work is based foremost on the principles of collaboration and teamwork. Achieving this requirement for astuteness is furthermore hindered by social and cultural factors in Vanuatu that implies doing things communally and for the common good rather than outshining others for personal benefit.

Bibliography

Clifford, J. (2001). Indigenous Articulations. *The Contemporary Pacific*, **13**(2): 467-90.

Conquergood, D. and Johnson, E. P. (2013). *Cultural Struggles: Performance, Ethnography, Praxis*. Ann Arbor: University of Michigan Press.

Dick, T. (2015). Chorographing the Vanuatu Aquapelago. *Shima: The International Journal of Research into Island Cultures*, **9**(2): 1-22.

Grossberg, L. (1986). On Postmodernism and Articulation: An Interview with Stuart Hall. *Journal of Communication Inquiry*, **10**(2): 45-60.

Grossberg, L. (1992). *We Gotta Get out of This Place: Popular Conservatism and Postmodern Culture*. New York: Routledge.

Hall, S. (1980). Race, Articulation, and Societies Structured in Dominance. *Sociological Theories: Race and Colonialism*. Ed. UNESCO. Paris: UNESCO.

Horst, H., Hjorth, L. and Tacchi, T. (2012). Rethinking Ethnography: An Introduction. *Media International Australia* 145 .

Madison, D. S. (2005). Performance Ethnography. *Critical Ethnography: Method, Ethics and Performance*. (Ed) Madison, D. Soyini. Thousand Oaks, CA: SAGE Publications, Inc., pp: 149-81.

Maxwell, I. (2012). Seas as Places. *Shima: The International Journal of Research into Island Cultures*, **6**(1): 27-29.

Olwig, K. R. (2008). Has 'Geography' Always Been Modern?: Choros, (Non)Representation, Performance, and the Landscape. *Environment and Planning* A, **40**(8): 1843-61.

Slack, J. D. (1996). The Theory and Method of Articulation in Cultural Studies. *Stuart Hall: Critical dialogues in cultural studies*, pp. 112-27.

Tawa, M. (2002). Place, Country, Chorography: Towards a Kinesthetic and Narrative Practice of Place. *Architectural Theory Review*, **7**(2): 45-58.

Sequentiality in Genetic Digital Scholarly Editions. Models for Encoding the Dynamics of the Writing Process

Wout Dillen
wout.dillen@uantwerpen.be
University of Antwerp, Belgium

When TEI P5 version 2.0 was published in 2011, scholarly editors who are interested in the writing process of literary works gained an important instrument for encoding their genetic Digital Scholarly Editions in TEI-conformant XML. After a long process of deliberation, this version of the TEI's encoding schema incorporated a large number of modifications proposed by the TEI MS SIG's Workgroup on Genetic Editions that aimed to re-evaluate the existing TEI tagset in order to facilitate the encoding of genetic phenomena (TEI Consortium, 2011). The Workgroup's 'Encoding Model for Genetic Editions' (2010) reveals two major points of interest in this proposal: (1) the need for the ability to encode features of the document rather than those of the text; and (2) the need for the ability to encode time, sequentiality and writing stages in those documents' transcriptions.

The main answer to the first point of interest was the introduction of the <sourceDoc> element (as well as its <surface> children), that was allowed to exist on the same hierarchical level as the <teiHeader>, <facsimile>, and <text> elements. Since the Text Encoding Initiative has (as its name implies) historically favoured 'text' over 'document', this can be regarded as a powerful statement to the TEI community that documents are as valuable as texts in textual scholarship, and that it should be possible to transcribe them as such. As a result, this encoding model has been gratefully adopted by editors who are taking a more document-oriented approach to the transcription of their materials – like those of the *Shelley Godwin Archive* (*S-GA*) for instance (Shelley, 2013). The question remains, however, whether the use of this vocabulary is enough to classify a Digital Scholarly Edition as a 'genetic' edition. While the document will take up a central position in any genetic edition, the use of the 'Genetic Editions' document-oriented transcription model is not a distinctive feature of the genetic edition in itself.

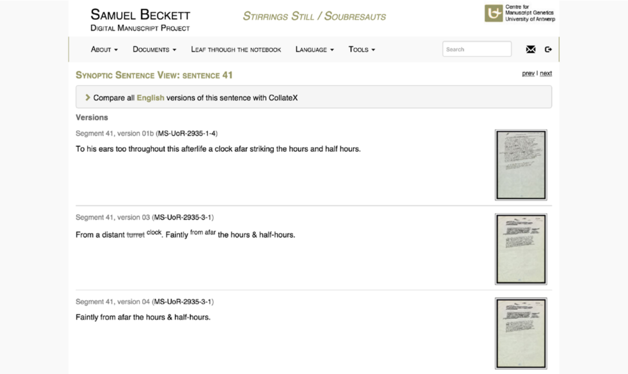
The Workgroup's second point of interest (the encoding of 'time') is much more central to genetic criticism. In 'The Open Space of the Draft Page', Daniel Ferrer makes a compelling argument that 'the draft is not a text […], it is a protocol for making a text', comparing it to a musical score that, though by itself inherently mute, can be interpreted as a set of instructions for a future performance (1998, 261). Likewise, a draft document leaves the writer with a set of instructions that help her transport the unfinished text from one writing stage to the next. The interpretation of these instructions, and of the distinction between different versions and writing stages, is one of the most important tasks of genetic criticism. This is what makes sequentiality such a key aspect of genetic editing: only by interpreting the draft materials as an interconnected sequence of writing acts can we expose the dynamics of the author's writing process.

There are many ways of encoding this sequentiality in the transcriptions of draft materials, across varying levels of granularity. The Workgroup's suggestion to use the <change> element to highlight distinct revision campaigns, for instance, effectively differentiates between individual versions of the same text when they are found within a single document. As Pierazzo and André's 'Proust Prototype' demonstrates, this method can even be employed to sequence individual stages within a single version (2012). Going even further, projects like the CD-ROM edition of Willem Elsschot's *Achter de Schermen* (2007) and the *Melville Electronic Library's* TextLab software (2009-) analyze what John Bryant has called the internal 'revision sequences' of individual sentences (Bryant, 2008). The danger of analyzing the writing process on this small a level, however, is that the *mechanics* of the writing process may start to interfere with the *dynamics* of that writing process. From a genetic perspective, it is more important to expose the dynamic relation between the textual elements involved in a modification (e.g. 'this is a substitution') than the mechanical order in which that modification was made (e.g. 'first this word was deleted, then this other word was added'). Since the exact writing sequence of such a modification is often impossible to reconstruct with any degree of certainty, consistently analyzing and sequencing all the work's revision sites may introduce a number of hypotheses in the edition that the editor is not necessarily comfortable with committing to.

On the other side of the spectrum, analyzing larger macrogenetic processes across documents, the 'Encoding Model for Genetic Editions' refers to the TEI's 'Graphs, Networks, and Trees' module, suggesting to encode the relations between documents as the <arc>s between <node>s in a <graph> element. Depending on the complexity of the writing process, this <graph> may result in an intricate data structure that can be used to visualize the chronology of the writing process on a highly abstract level. For writing processes that are less complex on the macrogenetic level, however, this model may be too much pain for too little gain, as a manually designed timetable could also do the trick.

The Beckett Digital Manuscript Project's approach to encoding sequentiality into its genetic Digital Scholarly Edition of Samuel Beckett's works tries to seek a middle ground between these two extremes: rather than analyzing the way in which individual sentences were written, the BDMP's encoding model allows the user to discover how those sentences were changed from version to version, across different documents. By linking related semantic clusters on the sentence level across versions, this model allows for the on the fly generation of a chronological overview of all the different versions of each sentence in the corpus. As such, this model combines the ability of comparing different versions of the same work of more macrogenetically oriented approaches with the higher granularity of more microgenetically oriented approaches.

After illustrating the challenges and opportunities of these different models of encoding sequentiality in genetic editions, this paper will demonstrate how the BDMP transcribes its genetic materials in view of visualizing their sequentiality in the edition's 'Synoptic Sentence View' (see 'Figure 1'). The paper will conclude by presenting an example of how this encoding model may also be used to interpret the macrogenetic writing sequence of individual documents by means of an animated visualization of the writing process of the first draft of Beckett's *L'Innommable*.


Samuel Beckett
Digital Manuscript Project
Stirrings Still / Soubresauts
Centre for Manuscript Genetics University of Antwerp
About
Documents
Leaf through the notebook
Language
Tools
Search
prev | next

Synoptic Sentence View: sentence 41

Compare all English versions of this sentence with CollateX

Versions

Segment 41, version 01b (MS-UoR-2935-1-4)

To his ears too throughout this afterlife a clock afar striking the hours and half hours.

Segment 41, version 03 (MS-UoR-2935-3-1)

From a distant ~~turret~~ $^{\text{clock}}$. Faintly $^{\text{from afar}}$ the hours & half-hours.

Segment 41, version 04 (MS-UoR-2935-3-1)

Faintly from afar the hours & half-hours.

Figure 1: BDMP Synoptic Sentence View

Bibliography

Beckett, S. (2011). *Stirrings Still / Soubresauts and Comment Dire / what is the word: a digial genetic edition (Series 'The Beckett Digital Manuscript Project' module 1)*, edited by Dirk Van Hulle and Vincent Neyt. Brussels and London, University Press Antwerp (ASP/UPA) and Bloomsbury Academic. http://www.beckettarchive.org (accessed on 12 November 2013).

Beckett S.(2013). *L'Innommable / The Unnamable: a digital genetic edition (Series 'The Beckett Digital Manuscript Project', module 2)*. Edited by Dirk Van Hulle, Shane Weller and Vincent Neyt. London and Brussels and London, Bloomsbury and University Press Antwerp (ASP/UPA) and Bloomsbury Academic. http://www.beckettarchive.org (accessed on 9 January 2014).

Beckett S. (2015). *Krapp's Last Tape / La Dernière Bande: a digital genetic edition (Series 'The Beckett Digital Manuscript Project' module 3)*, edited by Dirk Van Hulle and Vincent Neyt. Brussels and London, University Press Antwerp (ASP/UPA) and Bloomsbury Academic. http://www.beckettarchive.org (accessed on 22 October 2015).

Bryant, J.(2008). *Melville Unfolding. Sexuality, Politics, and the Versions of Typee: a Fluid-text Analysis, with an edition of the Typee manuscript*. Ann Arbor: University of Michigan Press.

Elsschot, W. (2007). *Achter de Schermen: elektronische editie* . Edited by Dirk Van Hulle, Vincent Neyt, and Peter de Bruijn. Kalmthout: Willem Elsschot Genootschap. CD-ROM.

Ferrer, D. (1998). The open space of the draft page: James Joyce and modern manuscripts. In *The Iconic Page in Manuscript, Print, and Digital Culture*, edited by George Bornstein and Theresa Tinkle. Ann Arbor: University of Michigan Press, pp: 249–67.

Melville, H. (2007). *Moby-Dick,* Edited by John Bryant and Haskell Springer, New York: Pearson / Longman.

Proust, M. (2012). *Autour d'une séquence et des notes du Cahier 46: enjeu du codage dans les brouillons de Proust–Around a Sequence and some Notes of Notebook 46: Encoding Issues about Proust's Drafts*. Edited by Elena Pierazzo and Julie André, with technical support from Raffaele Viglianti. London: King's College London. http://research.cch.kcl.ac.uk/proust_prototype/ (accessed 9 December 2015).

Shelley, M. (2013). Frankenstein. In *The Shelley-Godwin Archive*, directed by Elizabeth C. Denliger and Neil Fraistat. http://www.shelleygodwinarchive.org/ contents/frankenstein (accessed on 25 April 2015).

Workgroup on Genetic Editions. (2010). An Encoding Model for Genetic Editions. http://www.tei-c.org/Activities/Council/Working/tcw19.html (accessed on 30 October 2015).

TEI Consortium. (2011). P5 version 2.0 release notes. *TEI-C.* http://www.tei- c.org/release/doc/tei-p5-doc/readme-2.0.html (accessed on 13 November 2015).

correspSearch - A Web Service to Connect Diverse Scholarly Editions of Letters

Stefan Dumont
dumont@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Letters are an important historical source: First, they may contain comments from contemporaries about the most different events, persons, publications, and issues. Second, letters allow insights about connections and networks between correspondence partners. So, questions occur which can only be answered across the borders of scholarly letter editions due to the fact that these editions are usually focussed on partial correspondences (of a certain person or between two specific persons). But this requires time-consuming searches across various letter editions. This has been a well-known problem for quite some time, now (Bunzel, 2013: 117) and has already evoked work on a few databases dedicated to correspondence, like e.g. "Early Modern Letters Online".[1] But these databases have some limitations: firstly, they focus on specific research questions, time periods, geographic areas, or certain material. Secondly, they don't provide an open, standard-based and well-documented way to provide and update data. Furthermore, the data can be searched and displayed, but only on a simple website. In none of the existing databases dedicated to edited letters it is possible to query and retrieve the data via an Application Programming Interface (API) and under a free license for subsequent use.

The mentioned methodic problem has lead Wolfgang Bunzel, who works in the field of research about the Romanticism, to request:

> 'the creation of a decentralized, preferably open digital platform, based on HTML/XML and operating with minimal TEI standards, which is extensible in different directions and allows for existing web portals and websites to contribute at the lowest possible cost. This doesn't request some kind of super structure which covers the entire amount of letters from the Romantic era (which could not be estimated exactly, anyway) but rather an intelligent linking system, which associates existing documents with one another. The creation of such nexus will naturally lead to research options reaching from searches for persons and places to specific keyword-based searches […]' (Bunzel, 2013: 123)[2]

With "correspSearch" (http://correspSearch.bbaw.de) this paper will present a web service, which takes a step in this direction by aggregating metadata of letters from various (digital or printed) scholarly editions and providing them collectively via open interfaces (Fig. 1). In doing so, it is independent from specific research questions as well as from temporal, geographic, or thematic limits.

The basis of the web service are digital indexes of letters provided by the editions in the "Correspondence Metadata Interchange Format" (CMIF), that has been (and will further be) developed by the Correspondence Special Interest Group of the Text Encoding Initiative (TEI).[3] The CMI format is based on the TEI Guidelines and allows to interchange metadata of letters, postcards etc. between scholarly editions by restricting and normalizing the essential elements of a communication act, namely sender, addressee, dates and places of writing and receiving. Besides the consistent TEI XML encoding, interchange will be enabled by usage of ISO dates and authority controlled identifiers. Sender, addressee, sender's place as well addressee's place are identified unambiguously using authority IDs, as e.g. provided by the Library of Congress.[4] When reading the indexes of letters the web service retrieves the most common authority controlled IDs from the Virtual International Authority File (VIAF). This way, IDs from different authority files are mapped onto one another. Up till now, the web service supports VIAF, GND ("Gemeinsame Normdatei" from the Deutsche Nationalbibliothek) as well as the authority files of the Bibliothèque nationale de France (BNF), the Library of Congress (LC), and the National Diet Library (NDL) in Japan. As for place names the web service uses "GeoNames".

The scholarly editions themselves provide such digital indexes of letters in CMI format, online and under a free license (CC-BY 4.0). For this purpose the CMI format and its creation process is extensively documented on the correspSearch website including a FAQ section.[5] After providing a CMI file online, it is only necessary to register its URL for the web service. After that the file is automatically retrieved by correspSearch periodically (and in that way updated, if necessary).

The aggregated letter indexes are searchable on the correspSearch website by correspondent, location, and date. Correspondent and location can be specified according to their role in the communication process. Search results are displayed based on the metadata of the individual letter, together with biographical details. Letters from digital editions are directly linked.

Apart from the website an API has been implemented which allows for automatic requests to the web service.[6] In this scenario, the results are provided in TEI-XML in the described CMIF under a CC-BY 4.0 license, thus ensuring and facilitating further use and processing of the search results. Furthermore, the web service offers BEACON files as well as an experimental TEI-JSON output.

Thanks to the API it is possible to automatically refer or even link from one digital letter edition to related letters provided by other editions. This function was already implemented in a prototype for the digital scholarly edition "Schleiermacher in Berlin 1808-1834" (Fig. 2).[7] This feature helps researchers avoid methodological problems when interpreting a piece of correspondence: When analyzing a letter they usually consider the preceding and following letters in the correspondence between the sender and addressee, as well. However, their interpretation often does not include the letters which the correspondents send to or receive from *other* persons. With this feature the background of historical correspondences can be easily explored.

Via the API scholars can also exploit the data basis by usage of their own innovative technologies as well as of technologies which the web service itself does not yet support technically. Therefore, with a sufficiently extended data basis and the suitable software it will be possible to perform research on e.g. social or correspondence networks based on correspSearch.[8] Furthermore, the correspSearch API was connected with the web service "XTriples", developed by the Academy of Sciences, Humanities and Literature in Mainz (Germany). Thus the results can be converted into RDF and provided for further analyses with the help of semantic web technologies.[9]

The web service correspSearch and the CMIF are still under development. In the future it should be possible to search also for mentioned persons, events, publications etc. For this purpose the enhancement of the CMI file is currently discussed (Dumont, 2015). Also additional authority files will be supported, e.g. the Getty Thesaurus of Geographic Names.[10]

The web service correspSearch was granted the "Berlin Digital Humanities Award 2015" (First Prize, endowed with 1.200 €).

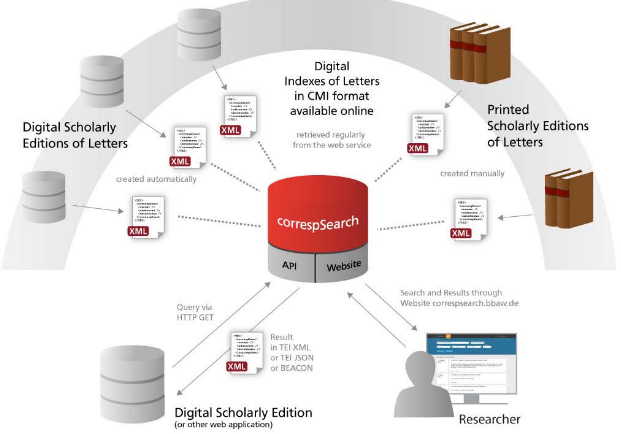

Digital Scholarly Editions of Letters
XML
created automatically
Digital Indexes of Letters in CMI format available online
retrieved regularly from the web service
Printed Scholarly Editions of Letters
created manually
correspSearch
API
Website
Query via HTTP GET
Result in TEI XML or TEI JSON or BEACON
Search and Results through Website correspsearch.bbaw.de
Digital Scholarly Edition
(or other web application)
Researcher

Fig 1: Operating principle of the web service correspSearch

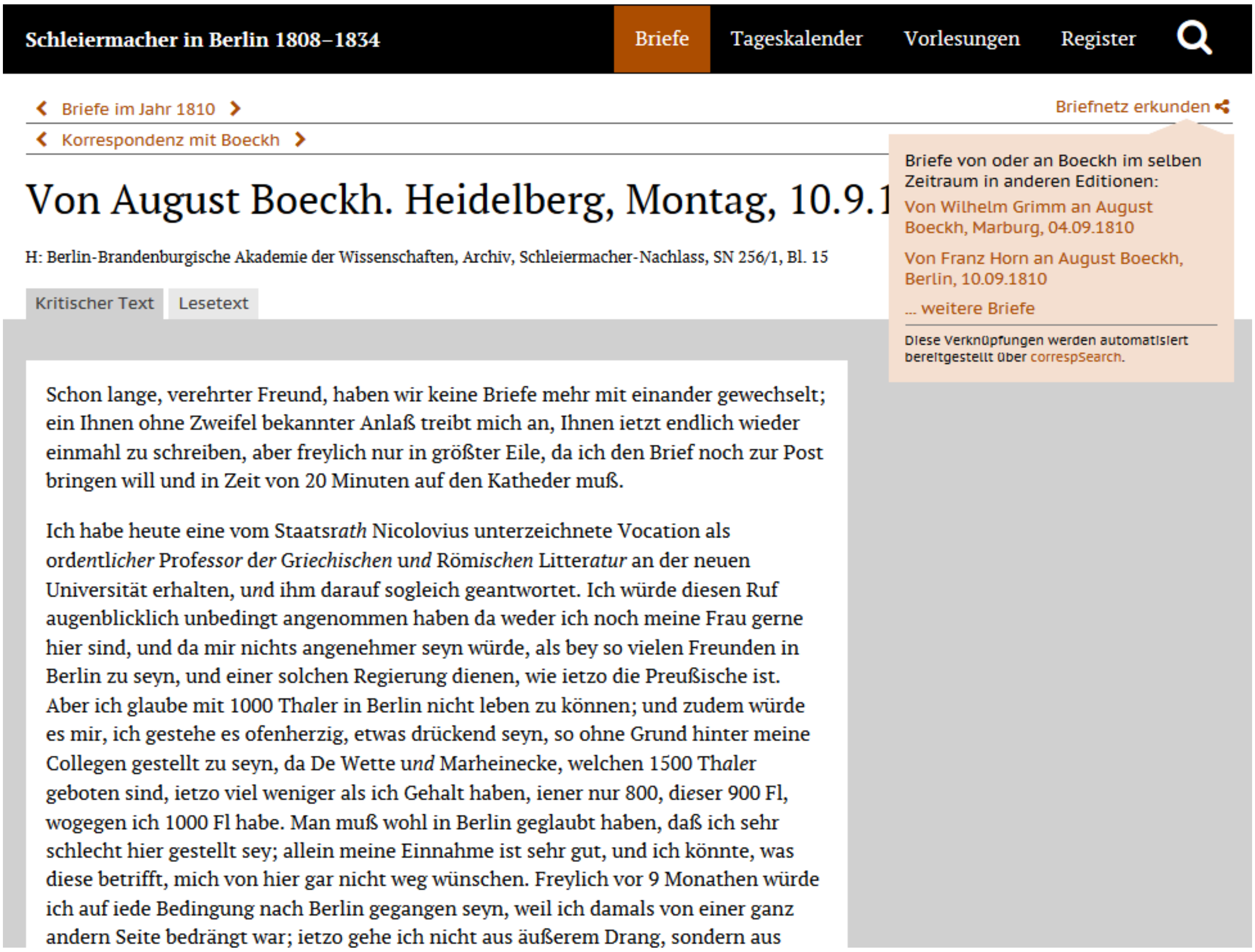
The digital scholarly edition queries the correspSearch API for other letters from and to August Boeckh around the date of the letter displayed (10 september 1810). If there is a result, the edition provides links to these letters.

Bibliography

Bunzel, W. (2013). Briefnetzwerke der Romantik. Theorie – Praxis – Edition. In Bohnenkamp, A. and Richter, E. (eds.), *Brief-Edition im digitalen Zeitalter* (=Beihefte zu editio Bd. 34). Berlin/Boston: de Gruyter 2013. pp. 109-31.

Dumont, S. (2015). Perspectives of the further development of the Correspondence Metadata Interchange Format (CMIF). *digiversity — Webmagazin für Informationstechnologie in den Geisteswissenschaften.*http://digiversity.net/2015/perspectives-of-the-further-development-of-the-correspondence-meta-data-interchange-format-cmif/ (accessed 27 February 2016)

Stadler, P. (2012). Normdateien in der Edition. *Editio*, **26**: 174-83.



Schleiermacher in Berlin 1808–1834

Von August Boeckh. Heidelberg, Montag, 10.9.1

H: Berlin-Brandenburgische Akademie der Wissenschaften, Archiv, Schleiermacher-Nachlass, SN 256/1, Bl. 15

Briefe von oder an Boeckh im selben Zeitraum in anderen Editionen:

Von Wilhelm Grimm an August Boeckh, Marburg, 04.09.1810

Von Franz Horn an August Boeckh, Berlin, 10.09.1810

... weitere Briefe

Diese Verknüpfungen werden automatisiert bereitgestellt über correspSearch.

Schon lange, verehrter Freund, haben wir keine Briefe mehr mit einander gewechselt; ein Ihnen ohne Zweifel bekannter Anlaß treibt mich an, Ihnen ietzt endlich wieder einmahl zu schreiben, aber freylich nur in größter Eile, da ich den Brief noch zur Post bringen will und in Zeit von 20 Minuten auf den Katheder muß.

Ich habe heute eine vom Staats*rath* Nicolovius unterzeichnete Vocation als ordent*licher* Professor der Griechischen *und* Römischen Litter*atur* an der neuen Universität erhalten, u*nd* ihm darauf sogleich geantwortet. Ich würde diesen Ruf augenblicklich unbedingt angenommen haben da weder ich noch meine Frau gerne hier sind, und da mir nichts angenehmer seyn würde, als bey so vielen Freunden in Berlin zu seyn, und einer solchen Regierung dienen, wie ietzo die Preußische ist. Aber ich glaube mit 1000 Thaler in Berlin nicht leben zu können; und zudem würde es mir, ich gestehe es ofenherzig, etwas drückend seyn, so ohne Grund hinter meine Collegen gestellt zu seyn, da De Wette u*nd* Marheinecke, welchen 1500 Thaler geboten sind, ietzo viel weniger als ich Gehalt haben, iener nur 800, dieser 900 Fl, wogegen ich 1000 Fl habe. Man muß wohl in Berlin geglaubt haben, daß ich sehr schlecht hier gestellt sey; allein meine Einnahme ist sehr gut, und ich könnte, was diese betrifft, mich von hier gar nicht weg wünschen. Freylich vor 9 Monathen würde ich auf iede Bedingung nach Berlin gegangen seyn, weil ich damals von einer ganz andern Seite bedrängt war; ietzo gehe ich nicht aus äußerem Drang, sondern aus

Fig 2: Screenshot of the digital scholarly edition "Schleiermacher in Berlin 1808-1834" (published soon), which presents letters to and from the theologian Friedrich Schleiermacher

Notes

[1] Early Modern Letters Online (EMLO), http://emlo.bodleian.ox.ac.uk/, is the one with the largest databases, but also includes data from other databases like "Circulation of Knowledge and

Learned Practices in the 17th-century Dutch Republic", http://ckcc.huygens.knaw.nl/. Besides this, there also exist more focused databases like "Exilnetz33", http://exilnetz33.de.

[2] Please note, that this is my own translation into English.

[3] The CMIF is maintained in a GitHub repository: https://github.com/TEI-Correspondence-SIG/CMIF

[4] For using authority files in scholarly editions, cf. Stadler 2012

[5] See http://correspsearch.bbaw.de/index.xql?id=participate

[6] See http://correspsearch.bbaw.de/index.xql?id=api

[7] The first version of the scholarly digital edition "Schleiermacher in Berlin" will be published in the next months by the Berlin-Brandenburg Academy of Sciences and Humanities.

[8] For example: the developers of the visualisation tool "nodegoat" imported the data in their application to visualize a correspondence network: http://correspsearch-test.nodegoat.net/viewer.p/4/136/scenario/1/geo/fullscreen

[9] http://xtriples.spatialhumanities.de. One prototype configuration is available under http://xtriples.spatialhumanities.de/examples.html

[10] http://www.getty.edu/research/tools/vocabularies/tgn/

Corpus Analyses of Multimodal Narrative: The Example of Graphic Novels

Alexander Dunst
dunst@mail.upb.de
University of Paderborn, Germany

Rita Hartel
rst@mail.upb.de
University of Paderborn, Germany

Sven Hohenstein
sven.hohenstein@uni-potsdam.de
University of Potsdam, Germany

Jochen Laubrock
laubrock@uni-potsdam.de
University of Potsdam, Germany

Introduction

This paper presents first empirical analyses and visualizations of a large corpus of graphic novels – an increasingly popular form of book-length comics aimed at adults – that is currently in the process of being assembled and digitized. We introduce an XML vocabulary and visual editor that we have developed for the annotation of our corpus, and reflect on the challenges presented by a cultural form that is characterized by the complex interaction of text and images. Analyzing the specific narrativity of this and other multimodal cultural forms (including illustrated books and magazines, theater, film, television and computer games), we argue, calls for a combination of quantitative and qualitative methods drawn from such diverse fields as narratology, digital art history, and cognitive science. In contrast to corpus analyses of literary texts, which have made great strides in recent years, comparable work on visual narrative still remains in its infancy and at the periphery of the digital humanities. While this can be traced, in part, to copyright issues, such scholarship also faces a number of crucial technical and methodological hurdles – from image description, classification, and object recognition to the operationalization of narratological concepts. Given the dominance of visual storytelling in modern and contemporary culture, overcoming these hurdles will represent an important contribution to the further development of DH research.

The introduction presents our corpus and the wider research questions of our interdisciplinary group. This is followed by a brief overview over the "Graphic Narrative Markup Language" (GNML), which builds on TEI, and the visual editor developed for the annotation of graphic novels, but which is also applicable to other multimodal forms. A version of this editor will be available as open-acess software by the time of the conference. Part two introduces a number of analyses and visualizations combining the study of text and images that make up graphic novels. The final part moves to the quantitative and qualitative analysis of graphic novels with the help of eye-tracking. This approach allows us to study the construction of storyworlds by empirical readers, and thus opens up an aspect of narrative that remains severely underrepresented in DH, and the humanities at large.

1. GNML-Editor: Tools for (Semi-)Automatic Annotation

Whereas the automatic analysis of text corpora has become feasible in many instances, such automation currently remains a pipe dream for multimodal narratives. In the case of comics and similarly hand-drawn, or otherwise non-perspectival, images, object identification depends on lengthy training efforts and registers a relatively high error rate. Similarly, standard OCR programs fail at recognizing the (quasi) hand writing that dominates comics. As a consequence, our corpus study presently depends on manual and semi-automatic annotation. For this purpose, we have developed the XML-language GNML, which builds on TEI and previous efforts by John Walsh (2012), to describe all textual and visual properties of graphic novels. To minimize errors during the annotation process, our visual GNML-editor supports annotators with integrated spell checking and auto-completion mechanisms. An automatic recognition of panels is complemented by a function that recognizes the borders of individual captions, speech bubbles, and characters to accelerate annotation. Further automations, such as an in-built OCR for narrative text that conforms to standard fonts, are currently under development. As the conceptual basis of the editor (visual objects with graphic and textual characteristics) is not limited to comics but can be applied to other text-image combinations in visual culture (from illustrated manuscripts to film and TV), the editor will be generalized for the annotation of such formats.

2. Quantitative Analyses and Visualizations of Graphic Novels

Part two presents approaches that combine image and text analysis for a number of structural features of graphic novels. Methods developed for digital literary studies, such as topic modeling, are of limited value for the analysis of visual culture given the dominance of images. In contrast, studies of large-scale image sets have so far shown little interest in narrative analysis. To complicate matters further, most narratological concepts are drawn from the study of literary texts, and it remains questionable to what extent they can be successfully applied to visual narrative.

In a first analysis of the corpus, which is still in the process of being digitized and annotated, we look at the historical development of the visual and textual elements of about 150 book covers of graphic novels. This includes a grammatical and semantic analysis of their titles with the help of a statistical language parser, as well as the stylistic and visual attributes of their design and cover images. In a second step, we move to more detailed studies of a first sub-corpus that consists of the ten most-cited titles within our larger set of graphic novels. Such a small sub-set does not allow for genre comparisons or for studying historical developments within the form. However, we can consider the narrative features of representative texts within our corpus. In order to do so, we compare a network analysis of characters with their visual prominence and respective share of text, and complement this with a stylistic analysis of the latter.

3. Eyetracking Analysis of Multimodal Narrative

The experimental observation of eye movements has proven a reliable measure of the human processing of text and images, and allows us to form hypotheses about the construction of storyworlds by empirical readers. The final part of the paper aims to show the value of this method by considering excerpts from a first, explorative corpus of canonical graphic novels. In contrast to theoretical scholarship on comics, which has emphasized the primacy of images (Groensteen, 2009), our experiments demonstrate that readers focus most of their attention on the text. Not only is it usually read first, but many images are either not focused on at all, or analyzed purely in peripheral vision. Whether images are viewed depends, among other variables, on their informational content: if either visual aspects or the storyline continue from one panel to the next, it is much more likely that a panel will be skipped by the reader than if they are distinguished more clearly from its immediate predecessor. We also look at the interaction between visual and textual levels: do reading habits differ if text and images refer to distinct storylines? Finally, we report on experiments that focus on comic reading expertise, for which we propose a new empirical measure. In sharp contrast to the reading of text alone, where experience and reading speed are positively correlated, experienced comics readers focus on the visual aspects of the panels for an extended amount of time. This time appears to be invested wisely, since they are able to better understand the story, as shown by an empirical content test. Taken together, these results suggest that the text and image work together to transmit the narrativity of graphic novels, and that a specific type of expertise is required to understand multimodal narratives. This maps well onto the hypothesis that comics and other forms of sequential art use a particular kind of visual language (McCloud, 1993), which has been analyzed in psycholinguistic terms by Cohn (2013).

4. Summary and Conclusions

We present a new DH project aiming at collecting and analyzing a corpus of graphic literature, enriched by human annotations as well as by a corpus of eye-movement recordings to measure the momentary distribution of readers' attention. First example analyses on both global and local levels demonstrate the potential of this approach. A toolchain for description, annotation, and analysis of these data is being developed, and is of potential use for a wider field of studies in cultural analytics of image-related and multimodal material. In perspective, the corpus will be further enhanced by automated description, using features developed in the field of computer vision (Farabet et al., 2013; Krizhevsky et al., 2012; Rigaud et al., 2015; Serre et al., 2007).

Bibliography

Cohn, N. (2013). *The Visual Language of Comics. Introduction to the Structure and Cognition of Sequential Images.* London: Bloomsbury.

Farabet, C., Couprie, C., Najman, L. and LeCun, Y. (2013). Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**: 1915–29.

Groensteen, T. (2007). *The System of Comics*. Jackson, MS: University of Mississippi Press.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**: 1097–1105.

Lowe, D. G. (2004). Distinctive Image Features from Scale-

Invariant Keypoints. *International Journal of Computer Vision*, **60**: 91–110.

McCloud, S. (1993). *Understanding Comics: The Invisible Art.* New York, NY: Harper Collins.

Rigaud, C., Guérin, C., Karatzas, D., Burie, J.-C. and Ogier, J.-M. (2015). Knowledge-driven understanding of images in comic books. *International Journal on Document Analysis and Recognition*, **18**: 199–221.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. and Poggio, T. (2007). Robust Object Recognition with Cortex-like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**: 411–26.

Walsh, J. (2012). Comic Book Markup Language: An Introduction and Rationale. *Digital Humanities Quarterly,* (6–1).

Déchiffrer Le Mythe De l'Amour

Michael Eberle-Sinatra
michael.eberle.sinatra@umontreal.ca
Université de Montréal, Canada

Marcello Vitali-Rosati
marcello.vitali.rosati@umontreal.ca
Université de Montréal, Canada

Les sites de rencontres ont acquis depuis plusieurs années une place centrale dans nos pratiques sociales. L'idée de pouvoir utiliser des dispositifs algorithmiques pour rencontrer des personnes n'est plus une nouveauté et n'est pas née avec Internet. Il suffit de penser qu'en France, plusieurs plateformes de rencontres étaient déjà disponibles sur Minitel dans les années 1980. Mais la rapide diffusion des connexions Internet et la naissance du web dans les années 1990 ont rendu possible un développement impressionnant de ce type de service. Match.com a été fondé en 1995 et est probablement aujourd'hui le plus utilisé au monde ; d'autres plateformes comme eHarmony ou OkCupid ou Plenty of Fish sont apparues dans les années suivantes, proposant chacune sa vision de la rencontre amoureuse.

Comment analyser ce phénomène? Comment le comprendre? Plusieurs approches sont possibles et beaucoup a été écrit sur le sujet. La question que nombre de sociologues et psychologues se sont posé est: est-ce possible de trouver l'amour sur Internet? Dans l'introduction du numéro spécial *Sociologies et Société*, Chiara Piazzesi note "la tension entre d'une part l'omniprésence discursive et sémantique de l'amour et d'autre part notre incapacité à savoir de quoi exactement il s'agit" (Piazzesi, 2014).

La question qu'il faut se poser par rapport aux sites de rencontre n'est donc pas tellement s'ils donnent lieu à des "vraies" rencontres, mais plutôt sur quel type de valeurs ils se basent et quelles valeurs ils produisent. En d'autres mots, il faut essayer de comprendre quelle idée de rencontre est proposée par ces services. Nous vivons de plus en plus dans une culture numérique: notre idée d'amitié est forcement conditionnée par la façon de concevoir l'amitié de Facebook, notre idée de pertinence est façonnée par Google, notre rapport à l'espace et au temps structuré par les dispositifs de vidéoconférence, les GPS, etc. Il est certain que notre idée d'amour est ou sera elle-aussi influencée par nos pratiques numériques.

Or les plateforme existantes proposent des types de rencontres très différentes : certaines promettent de trouver rapidement des relations sexuelles (comme AdultFriendFinder, par exemple), d'autres promettent de trouver la personne avec qui se marier (eHarmony). Nous nous concentrons sur les sites qui sont explicitement axés sur l'idée d'amour : ce que ces plateformes visent est de rendre possible une rencontre de laquelle pourra naître une relation amoureuse. Deux exemples pertinents sont Match.com et OkCupid. Concrètement ces sites permettent la création d'un profil et mettent en place un algorithme qui relie les profils entre eux : comme tous les autres algorithmes de recommandation, l'algorithme des sites de rencontre est un ensemble de règles formelles qui permettent l'analyse des données d'un profil pour le mettre en relation avec un autre profil.

La question qu'il semble fondamental de poser est donc quelle conception de l'amour se cache derrière ces règles? Qu'est-ce que l'amour pour une plateforme comme OkCupid? Quelle est l'idée d'amour à partir de laquelle sont pensées les règles formelles qui constituent l'algorithme? Cette question s'accompagne forcement de la question opposée; car si d'une part les algorithmes se basent sur une idée de ce qu'est l'amour, en même temps, dans la pratique, ils ont un effet normatif: en d'autres termes ils produisent eux-mêmes une idée d'amour.

Commençons par une note méthodologique: il est nécessaire de préciser que nous n'avons bien évidemment pas accès au code source de ces algorithmes. L'ensemble du code est propriétaire et il n'est donc pas possible de l'analyser d'un point de vue mathématique ou informatique. On est dans la même situation quand on essaye de comprendre le fonctionnement de PageRank. Nous n'avons accès qu'aux textes avec lesquels les entreprises communiquent sur leur algorithme. Si cette limitation empêche une évaluation objective du fonctionnement des algorithmes, elle ne limite pas tellement le type de questionnement que nous proposons ici : il ne s'agit pas, en effet, de comprendre le réel fonctionnement des algorithmes mais d'analyser leurs bases culturelles, leurs valeurs. La communication publicitaire sur leur fonctionnement est donc déjà un excellent point de départ.

Les algorithmes peuvent facilement donner une réponse à la première: la plateforme va chercher dans les profils de millions de personnes et avec sa capacité de calcul l'algorithme sera capable de trouver la bonne personne, la seule bonne personne, au milieu d'un nombre immense de profils. La rencontre est donc nécessaire : les deux personnes qui se rencontrent ont été sélectionnées dans la totalité – ou presque – des personnes possibles. La puissance de calcul assume en quelque sorte le rôle du destin: grâce à la capacité de regarder dans une énorme masse de données, on sera capable d'aller au delà des aléas qui pourraient empêcher une rencontre. Or dans les faits cette idée n'est pas du tout respectée par la pratique de ces plateforme, car les rencontres proposées sont très ciblées et, par exemple, très axées sur la proximité géographique. Mais la promesse est d'aller chercher dans l'ensemble des profils. Cette promesse permet de ne pas assimiler la nature du service offert à une pure démarche commerciale. Il ne s'agit pas de « vendre » le bon profil, mais d'identifier la bonne personne, la seule avec qui une relation amoureuse sera ensuite possible. Ce type de rhétorique caractérise la quasi-totalité de ces plateformes. Une analyse attentive de ce discours nous fait comprendre que ce que les sites de rencontres essayent de proposer – du moins dans leur discours – est vraiment une relation amoureuse, une *romance*.

Mais la puissance de calcul pourrait être un obstacle à la deuxième idée que nous venons d'évoquer: comment retrouver, dans le cadre d'une rencontre "calculée" par une machine la magie qui devrait caractériser l'amour? Comment ces plateformes peuvent-elles rendre possible le coup de foudre qui caractérise, dans notre imaginaire la relation amoureuse? Deux réponses sont possibles. D'une part on peut constater que plusieurs sites de rencontres essayent de réintroduire un élément de hasard ou de non calculé dans leurs propositions. C'est notamment l'exemple d'OkCupid qui proposa durant une période des "blind dates". L'algorithme met en relation deux personnes de façon aléatoire et propose un rendez-vous sans dévoiler le profil de la personne qu'on va rencontrer.

Mais il y a un autre facteur qui peut relier l'amour calculé à l'amour magique: la complexité du dispositif technologique est très souvent perçue comme magique. Le fait qu'on ne connait pas les algorithmes et que leur fonctionnement est mystérieux pour la plupart des usagers implique une sorte de rapport magique à l'objet technologique. Le fait qu'il y ait un calcul n'enlève donc rien à la possibilité de retrouver un côté de hasard et de magie.

De cette manière, les algorithmes essayent de ne pas mettre en danger l'aspect d'enchantement qui semble devoir caractériser une relation amoureuse. Cela pousse à penser qu'au lieu que transformer l'amour en une marchandise, ces plateformes essayent de récupérer un élément d'authenticité: en d'autres termes on pourrait affirmer que les sites de rencontres ont comme effet de promouvoir même auprès des sceptiques l'idée de la possibilité d'une rencontre amoureuse romantique. L'élément algorithmique se présente comme une sorte de garantie rationnelle de la possibilité de l'amour. Il nous semble que trois conceptions de l'amour influencent la notion qui se cache derrière les sites de rencontres : celle du Moyen-Âge, l'idée romantique et la conception du cinéma hollywoodien. Au Moyen- Âge on peut notamment retrouver le rapport étroit entre amour et vision qui semble être un des piliers de l'amour proposé par les sites de rencontres : on peut tomber amoureux car on peut tout voir. Dans l'amour romantique et surtout dans une réinterprétation *mainstream* hollywoodienne de l'amour romantique, on peut retrouver l'idée de la nécessité de la rencontre et celle de la magie. Une analyse de ces topoï littéraires accompagnée par une étude approfondie des textes qui présentent le fonctionnement des algorithmes pourrait nous aider à mieux cerner la façon qu'ont les sites de rencontres de concevoir l'amour. L'impact du numérique est culturel et les pratiques numériques façonnent nos visions du monde. Comprendre quelle est l'idée d'amour qui s'exprime à travers les algorithmes des sites de rencontres devient donc indispensable pour comprendre quelle est la conception de l'amour dans notre société.

Bibliography

Cancian, F. and S. Gordon. (1988). Changing Emotion Norms in Marriage: Love and Anger in U.S. Women's Magazines since 1900. *Gender and Society*,**2**(3): 308-42.

Cardon, D. (2013). Dans l'esprit du PageRank : Une enquête sur l'algorithme de Google, *Réseaux* 1, pp. 63-95.

Evans, M. (2002). *Love, an Unromantic Discussion*. Cambridge, Polity Press.

Giddens, A. (1992). *La transformation de l'intimité. Sexualité, amour et érotisme dans les sociétés modernes*. Rodex, Le Rouergue/Chambon.

Henchoz, C. (2014). La production quotidienne de l'amour en Suisse et au Québec : compatibilités intimes. *Sociologies et sociétés*, **46**(1): 17-36.

Piazzesi, Ch. (2014). Tout sauf l'amour ou porter un regard sociologique sur l'intimité amoureuse. *Sociologies et sociétés,* **46**(1): 5-14.

Slater, D. (2013). *Love in the Time of Algorithms: What Technology Does to Meeting and Mating*. Penguin Group.

Historical Linguistics' New Toys, or Stylometry Applied to the Study of Language Change

Maciej Eder
maciejeder@gmail.com
[1] Institute of Polish Language, Polish Academy of Sciences;
[2] Pedagogical University, Krakow

Rafał Górski
rafalg@ijp-pan.krakow.pl
[1] Institute of Polish Language, Polish Academy of Sciences;
[2] Jagiellonian University, Krakow

Background

In the last decades, quantitative linguistics (following exact and social sciences) has developed several statistical methods providing an insight into measurable phenomena of natural language. Although to a lesser extent, it also applies to the analysis of diachronic changes. Obviously, the so-called philological method in historical linguistics (unlike historical comparative and internal reconstruction methods) was always a kind of "corpus linguistics", which means that a linguist studying a given period of a language investigated, via close reading, available written records. Consequently, the text was usually treated as a (mistrustful) informant. The implication of this attitude is that in principle, a single attestation of a linguistic fact in a text was considered a strong evidence. Paradoxically, it is synchronic corpus linguistics that changed the overly conservative approaches to diachrony. The most significant here is the shift from purely qualitative to quantitative argumentation. Certainly, the availability of machine-readable corpora allows for much more sophisticated quantitative analysis these days.

A significant drawback of many of the quantitative methods applied so far is a tacit assumption that the researcher knows in advance which elements of the language are subject to change. In other words: the method of, say, plotting and inspecting the trend for a given phenomenon may be applied only to verify hypotheses stipulated earlier by traditional (that is qualitatively oriented) diachronic linguistics. A real challenge, however, is to develop such a method that would allow to trace chronological change in the language without a prior knowledge which linguistic features are responsible for the change. Promising results may be expected using some of the stylometric techniques based on the statistical analysis of style, especially the so-called multidimensional methods. The combination of stylometry and historical linguistics is not an entirely new idea. The problem of automatic recognition of relative chronology of texts was recently addressed by Stamou (2008; 2009), Štajner and Mitkov (2011), Popescu and Strapparava (2013), Štajner and Zampieri (2013), Zampieri et al. (2015). We shall note, however, that the first who sought to solve the question of chronology of texts via their stylistic features was Lutosławski (1897).

Stylometric methods are particularly efficient when applied to frequencies of function words (or, the most frequent words). However, an interesting question arises what if we disregard words and examine grammatical features instead? Obviously, the usage of archaic vs. modern inflected forms alone will differentiate texts written in two distinct (yet still close) periods. What is less obvious, however, is whether processing solely POS-tags, i.e. grammatical labels, can show the dynamics of language change. Note that the sequences of POS-tags are a good approximation of syntax, even if they cannot replace parsing (Hirst and Feiguina, 2007; Wiersma et al., 2011). To scrutinize the above research question, we performed a number of stylometric tests using different (tailored) methods and different combinations of lexical and grammatical features' *n*-grams.

Chronology at a glance

Standard stylometric methods are aimed at tracing differences between (groups of) texts. They proved to be successful in detecting a predominant stylistic signal, which in most cases is the authorial voice. However, when the number of analyzed texts is high enough, the emerging authorial groups (clusters) tend to form larger lumps reflecting the existence of other stylometric signals, such as genre, gender or chronology. This phenomenon can be observed very clearly when bootstrap consensus network – an enhanced version of cluster analysis (Eder, 2015) – is applied.

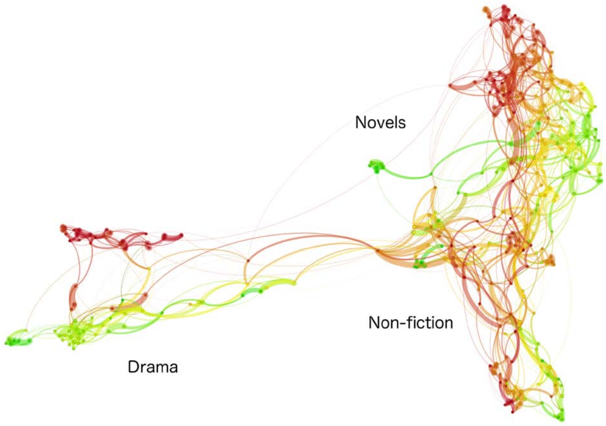


Figure 1: Stylometric network of similarities between 333 English texts

In Fig. 1, a network of an exemplary corpus of 333 English texts (De Smet, 2005) covering the period 1700–1930 is shown. The network was produced using most frequent words as predictors. One can notice a clear split into three distinct areas of the network that is due to a strong genre signal. However, despite the overwhelming division into novels, non-fiction, and drama, an additional chronological signal – represented by a transition from green (the earliest works) to red (the latest works) is fairly noticeable within each of the three sub-groups. Networks for other style markers (word *n*-grams, POS-tag *n*-grams) showed a similar behavior.

Modeling stylistic drifts

Certainly, the general picture revealed by the above network is by no means satisfactory, at least from the perspective of historical linguists. In particular, one would like to know how to pinpoint the observed chronological transition, in terms of identifying interpretable trends and/or breaks. The idea discussed in this section addresses this problem by combining multivariate stylometry with linear regression models.

Multidimensional scaling is a way of compressing (or projecting) a highly complex space into its simpler, usually two-dimensional, representation. Even if such a procedure always involves some loss of information, it is believed to reveal actual differences between samples. Now, since the technique allows to reduce the original space into an arbitrary number of dimensions, one can squeeze the data into just one dimension. This single MDS score can be plotted against the timeline, in order to test if any correlation between the two variables exist. The more diagonal is the shape of the plotted points, the higher the correlation.

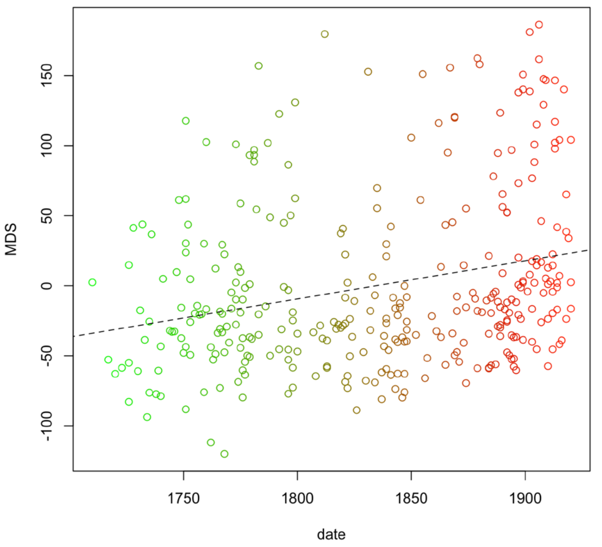


Figure 2: Multidimensional Scaling of 333 English texts (250 most frequent word 3-grams), compressed into one dimension and plotted against the timeline

In Fig. 2, some correlations between the timeline and the MDS values are fairly visible with a naked eye. However, when the results are scrutinized using a standard linear regression model $yi = xi\beta_1 + \beta_0 + \varepsilon$ (where β_1 and β_0 are parameters of the model, and ε is a random effect), their correlations become even more obvious. The estimated model (a dashed line in Fig. 2), is formulized as $\hat{y}_i = 0.272 * d_i - 499.94 + \varepsilon$, where *d*i denotes the *i*-th text's date of publication. In terms of the *p* value, the model is statistically significant ($p < 0.01$); however, the goodness of fit as represented by the adjusted *R*2 value is rather poor (*R*2 = 0.06), due to the overwhelming genre signal hidden in the dataset. When one splits the corpus into three genres and analyzes them separately, however, the explanatory power of the model is far higher.

Supervised classification and the timeline

One of the most interesting aspects of language development – overlooked in a vast majority of the existing studies – is the question of the dynamics of linguistic changes. Presumably, one should expect epochs of substantial stylistic drift followed by periods of stagnation, rather than purely linear trends.

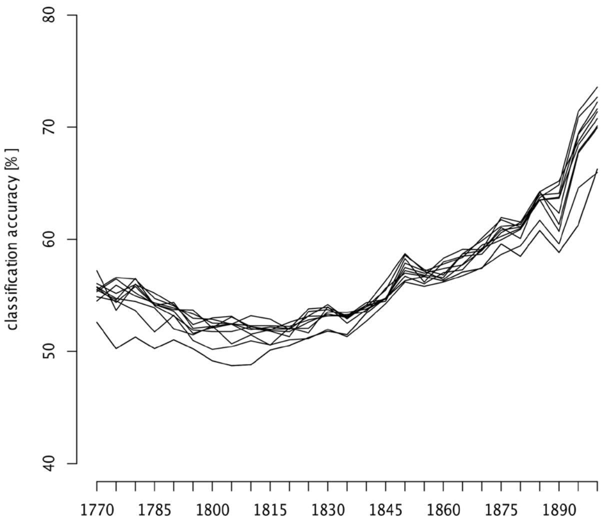


Figure 3: A sequence of Nearest Shrunken Classification tests on 333 English texts: cross-validated results for different vectors of most frequent POS-tag 2-grams

To assess this issue, we apply an iterative procedure of automatic text classification. First, we formulate a working hypothesis that a certain year – be it 1750 – marks a major linguistic break. We divide the text samples into the *ante* and *post* subsets, according to particular texts' publication date. Next, we randomly pick a number of train and test samples representing the both classes (*ante* and *post*), and we train a supervised classifier. We perform a standard classification, and record the cross-validated accuracy rates. Then we dismiss the original hypothesis, in order to test new ones: we iterate over the timeline, testing the years 1755, 1760, 1765, 1770, ... for their discriminating power. The assumption is simple here: any acceleration of linguistic change will be reflected by higher accuracy scores.

In Fig. 3, the classification accuracy rates for the aforementioned corpus of 333 English texts were shown (POS-tag 2-grams, NSC classifier). As one can observe, the scores obtained for the period 1750–1850 are only slightly higher than the baseline, betraying no revolutionary changes in this period. Later, however, the stylistic drift accelerates, reaching 70% of correctly recognized test samples.

Conclusions

in this paper we used a set of tailored stylometric methods to assess the question of language change over time. Our chosen techniques proved to be useful indeed; the further research will focus on tracing the very linguistic features that were responsible for the observed change. However, an important question has to be asked here: is it a change of Saussurean *langue* what we track with our approach, or rather the change of *parole*. Obviously, if texts written earlier can be separated from texts written more recently, they must share some features common for a given stage of language development. However, it is not clear if an observed change is due to, say, literary taste of the epoch or, if we face an actual change in the system here. Theoretically, the former and the latter are possible, as well as both answers together. It is also very likely that the change takes place in between: in the *norm* in the sense proposed by Coseriu (1958). Still there are no means to answer this question with any stylometric method, what for a linguist might be seen as a drawback. However, the proposed method informs the linguist about the fact of change, which takes place not only in lexis but also in syntax; about the speed of change and, above all, about the points where this speed accelerates.

Acknowledgements

This research is part of project UMO-2013/11/B/HS2/02795, supported by Poland's National Science Centre.

Bibliography

De Smet, H. (2005). A corpus of Late Modern English texts. *International Computer Archive of Modern and Medieval English*, **29**: 69–82.

Eder, M. (2015). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, **30**, first published online 3 December 2015, doi: 10.1093/llc/fqv061.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, **22**(4): 405–17.

Lutosławski, W. (1897). *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of His Writings*. London: Longmans, Green and Company.

Popescu, O. and Strapparava, C. (2013). Behind the times: Detecting epoch changes using large corpora. *International Joint Conference on Natural Language Processing*. pp. 347–55 http://anthology.aclweb.org/I/I13/I13-1040.pdf (accessed 25 November 2015).

Štajner, S. and Mitkov, R. (2011). Diachronic stylistic changes in British and American varieties of 20th century written English language. *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage at RANLP*. pp. 78–85 http://www.anthology.aclweb.org/W/W11/W11-41.pdf#page=88 (accessed 25 November 2015).

Štajner, S. and Zampieri, M. (2013). Stylistic changes for temporal text classification. *Text, Speech, and Dialogue*. Springer, pp. 519–26 http://link.springer.com/chapter/10.1007/978-3-642-40585-3_65 (accessed 25 November 2015).

Stamou, C. (2008). Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, **23**(2): 181–99.

Stamou, C. (2009). *Dating Victorians*. VDM Publishing.

Wiersma, W., Nerbonne, J. and Lauttamus, T. (2011). Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, **26**(1): 107–24.

Zampieri, M., Ciobanu, A. M., Niculae, V. and Dinu, L. P. (2015). AMBRA: A Ranking Approach to Temporal Text Classification. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, pp. 851–55 http://www.researchgate.net/profile/Marcos_Zampieri/publication/273769693_AMBRA_A_Ranking_Approach_to_Temporal_Text_Classification/links/550c14850cf2b2450b4e901e.pdf (accessed 25 November 2015).

Go Set A Watchman while we Kill the Mockingbird in Cold Blood, with Cats and Other People

Maciej Eder
maciejeder@gmail.com
[1] Institute of Polish Language, Polish Academy of Sciences;
[2] Pedagogical University, Krakow, Poland;

Jan Rybicki
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland

Even before Harper Lee's "new" book, *Go Set A Watchman*, was published earlier this year (2015), rumors as to its authorship abounded. Alabama police looked into alleged abuse of Lee's rights; suspicion suddenly (re) surfaced about the strange fact that one of the greatest bestsellers in American history was its author's only completed work; Lee's childhood friendship with Truman Capote (portrayed as Dill in *To Kill A Mockingbird*) and their later association on the occasion of *In Cold Blood* fueled more speculations on the two Southern writers' possible, or even just plausible, collaboration; finally, the role of Tay Hohoff, Lee's editor on her bestseller, was discussed. Desperate media turned to the usual front for stylometry, Matt Jockers, who graciously ceded this opportunity onto us. A story about our early results appeared in *The Wall Street Journal* (Gamerman, 2015), and it echoed even in our native Poland, where the country's major newspaper, *Gazeta Wyborcza*, also devoted a whole page to this international success of Polish stylometry (Makarenko, 2015).

The truth proved to be at once much less sensational than most of the rumors – and much more interesting. Stylometric evidence is very strong in this case: Harper Lee is the author of both *To Kill A Mockingbird* and *Go Set A Watchman*. The first method applied here was part of stylo, a stylometric package (Eder et al., 2013) for R (R Core Team, 2014): series of most-frequent word frequencies in a collection of texts were compared using Burrows's Delta measure of distance (Burrows, 2002); Delta distances were compared for each pair of the texts in this corpus by cluster analysis, and the results of clustering were used to create a bootstrap consensus tree. The resulting Fig. 1 shows the two Harper Lee books as two nearest neighbors just as it does the other authors included for comparison here. More importantly, perhaps, Truman Capote is far away. Most importantly, her editor's only available book, *Cats and Other People*, betrays no similarity to her charge. Since this sort of diagram is oriented at deciphering the strongest signal in word usage, authorship, the various rumors should be finally set at rest – the more so as the two Harper Lee novels have always been each other's nearest neighbors in a whole series of rigorous machine-learning classification tests performed using stylo's "classify" function.

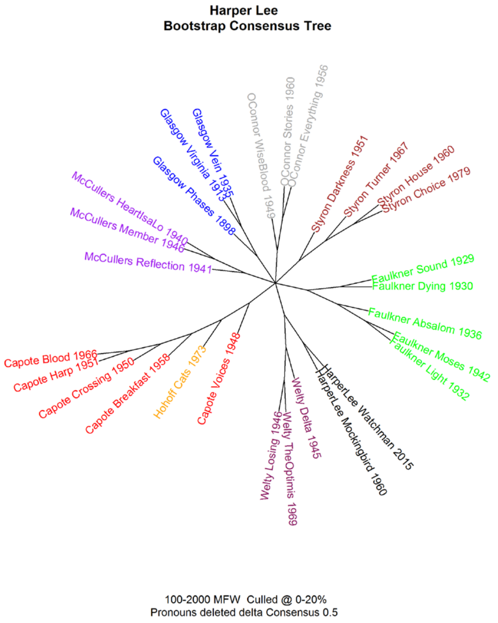


Figure 1: Harper Lee and selected authors of the American South, compared at 100–2000 most frequent words

Lesser affinities between texts are preserved in Fig. 2, which presents a network analysis of the same data treated with an enhanced version of the aforementioned consensus statistical method (Eder, 2015b) and produced with the Force Atlas 2 layout (Jacomy et al., 2014) in Gephi (Bastian et al., 2009). The degree of similarity is shown by the thickness of the curves that connect the particular texts: the thicker the line, the stronger the similarity. Additionally, the algorithm also spatially distributes the nodes (representing each text) to provide an additional visualization effect.

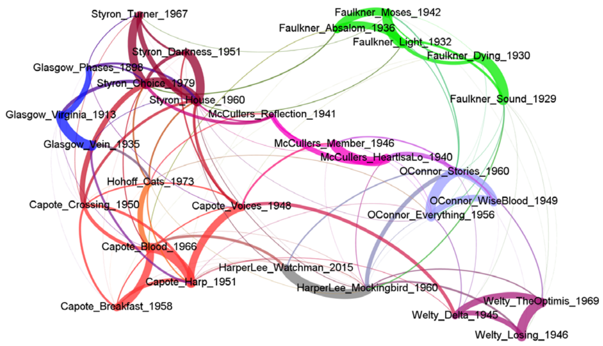


Figure 2: Network analysis of the same collection of novels

It is no surprise that this diagram echoes the previous one as far as the strongest similarities are concerned. Lee is still Lee; now, Faulkner stands almost alone. But then the lesser forces, represented by the slightly narrower connections, also count. The first thing that strikes the eye in the Lee neighborhood is the *Watchman*'s affinity to *In Cold Blood* and a more heterogeneous pattern for the *Mockingbird*: the book researched by Capote with Lee is still linked to her 1960 bestseller, but now only by the minutest of lines. This rephrases the Lee/Capote question in a more interesting way. Is there a drop of Capote in Lee? Perhaps not in the entirety of her work – perhaps just in a passage or two. This should be answered with a modification of the method: since it is difficult to see overlapping stylometric signals in an entire novel, one can see much more when the novel is split into equal and smaller fragments; then, the usual stylometric analysis is applied to the particular slices according to the "rolling. classify" procedure (Eder, 2015a).

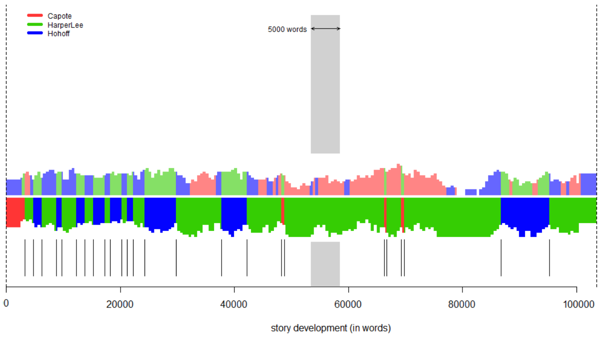


Figure 3: To Kill a Mockingbird contrasted sequentially against Capote's In Cold Blood (red), Hohoff's Cats and Other People (blue) and Lee's Go Set A Watchman (green). The lower band represents the strongest authorial signal; the upper band (in less intense colors) is the second-strongest signal

The most reasonable texts to be thus compared to *To Kill A Mockingbird* are Capote's *In Cold Blood* (since Lee helped with the research for that book), Lee's own *Go Set A Watchman* (to see how much of the *Watchman* might be found in the *Mockingbird*) and Tay Hohoff's *Cats and Other People* (to find out how much of Lee's rewriting of her original proposal might have been influenced by her experienced editor). This is presented in Fig. 3, and the result is quite interesting.

The signal in a little more than a half of the segments in *To Kill A Mockingbird* is that of the novel she originally brought to be published by Lippincott. It is highly significant that its longest stretch coincides with the trial that was only mentioned in the *Watchman* and became the focus of the book in the *Mockingbird*. This seems to suggest that while this refocusing of the book was made following the advice of the editor, the rewriting was indeed done by Harper Lee.

The rest of the *Mockingbird* is a veritable mosaic of her own and her editor's hand. Tay Hohoff's impact seems to be especially visible towards the end of the story, and it coincides with the novel's climax in Chapter 28: Scout, dressed in her elaborate and cumbersome ham costume, is attacked by Bob Ewell, who, following the struggle with Jem and then with Arthur "Boo" Radley, is left with his own knife stuck under his ribs.

We will never know, of course, whether Tay Hohoff really wrote that scene (and the others that seem to bear her mark) for Lee. But it is sensible to argue that while *To Kill A Mockingbird* is obviously a novel by Harper Lee, traces of someone who helped her along the way for two whole years – and who, at one point, talked the author into running down to the street to collect the manuscript that had been flung through the window in frustration (Shields, 2006: 121) – must be there somewhere. The results produced by the different functions of stylo are not in conflict when they show the overall strength of the *Watchman* signal in the *Mockingbird* and the possible echoes of Hohoff (or even, at the very onset of the novel, of Capote) in selected segments. Rather, they seem to provide new insights into the traces of various people involved in the making of a novel – and into how some of these traces may be identified and discerned by stylometry. It is equally sensible to find such traces in a work of a very particular kind: a novel that has been reprocessed almost beyond recognition in a long process of authorial and editorial collaboration; where the final version keeps the setting and the characters of the first, but changes its focus, its historical moment in time and, perhaps more importantly, its ideological message.

Bibliography

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Burrows, J. (2002). "Delta": A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, **17**: 267–87.

Eder, M. (2015a). Rolling stylometry. *Digital Scholarship in the Humanities*, **30**, first published online: 7 April 2015, doi: 10.1093/llc/gqv010.

Eder, M. (2015b). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, **30**, first published online 3 December 2015, doi: 10.1093/llc/fqv061.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, **8**, first published online 30 December 2015, https://journal.r-project.org/archive/.

Gamerman, E. (2015). Data Miners Dig Into "Watchman". *The Wall Street Journal*, 17 July 2015: D5, http://www.wsj.com/articles/data-miners-dig-into-go-set-a-watchman-1437096631.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, **9**(6): e98679, doi:10.1371/journal.pone.0098679.

Makarenko, V. (2015). Literackie śledztwa Polaków. *Gazeta Wyborcza*, 31 July 2015: 18.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Wien, http://www.R-project.org/.

Shields, C. J. (2006). *Mockingbird: A Portrait of Harper Lee*. New York: Henry Holt and Co.

OVAL: A Virtual Ecosystem for Immersive Scholarship and Teaching

Bill Endres
bill.endres@ou.edu
University of Oklahoma, United States of America

Matthew Cook
mncook@ou.edu
University of Oklahoma, United States of America

Will Kurlinkus
wkurlinkus@gmail.com
University of Oklahoma, United States of America

Introduction

3D modeling is transforming scholarship and teaching. In 2004, when Taliban militia destroyed the Buddhas of Bamiyan (one, the tallest Buddha in the world), a team of researchers from the Swiss Federal Institute of Technology responded by generating 3D models of these lost Buddhas (Grün et al., 2004). Indeed, 3D and virtual reality (VR) have sparked scholarly interest because VR projects, such as the Buddhas and Rome Reborn (an international effort to generate a virtual surrogate of the ancient city), allow for the modeling of inaccessible or lost historic spaces and objects and the interactive testing of scholarly hypotheses (Dylla et al., 2010).

Now, with advances in digital imaging and replication, 3D models have proliferated. They can be easily downloaded from sites such as SketchFab, Smithsonian X3D, and NASA 3D Resources. Moreover, scholars can publish them in a new journal: *Digital Applications in Archaeology and Cultural Heritage*. However, this proliferation raises questions about best practices for viewing, archiving, and interacting with these complex digital assets. The primary interface for 3D remains the flat, relatively small, computer screen—miniscule when compared to a 53-meter Buddha.

One solution is to provide an immersive experience through VR, enabling viewers to share space with 3D artifacts. Such an interface, however, presents further problems. VR systems tend to follow one of two models: 1. The model of Rome Reborn, in which a virtual system houses one surrogate of a physical space, or 2. A laboratory, such as Stanford University's CAVE system, in which sophisticated and costly VR equipment is employed for research. Can a virtual ecosystem be developed that not only makes uploading, hosting, and viewing VR assets easy but also is cost efficient and yet provides a robust, flexible, and accessible environment to meet divergent needs across a university and beyond?

In this presentation, we will discuss our answers to these questions derived from building a VR ecosystem housed in the library of the University of Oklahoma: the Oklahoma Virtual Academic Laboratory (OVAL). OVAL is a fully functioning VR environment. It is designed to deliver ease in uploading and engaging 3D assets, especially for non-technical users. Through digital immersion, multiple scholars and students simultaneously encounter VR assets, moving through and around and rotating and resizing them. Such engagement represents a paradigmatic shift in the computer interface: scholars and students are no longer limited by a flat monitor and instead share the virtual space with their objects of study.

Furthermore, OVAL encourages enhanced experiential learning (Hermon and Kalisperis, 2011; 60-61). It gives students and scholars opportunities to engage digital materials beyond their normal reach (such as interacting with atomic structures or cultural heritage artifacts). And, when partnered with museums, medical centers and other stakeholders, OVAL makes VR accessible to a wide scope of people, institutions, and industries. Thereby, OVAL offers a new model of accessible VR.

Challenges

Building a VR ecosystem for a whole university poses daunting challenges. For OVAL, the overriding hurdle was cost. To accommodate scholars and students across the disciplines, building multiple, multi-million dollar VR labs was not feasible. Therefore, cost constantly reigned-in decisions but rarely reigned-in performance.

There were also challenges for choosing compatible hardware and software. To keep costs reasonable and minimize spatial requirements, we opted to have users seated while immersed in OVAL, with head, upper body, and hands tracked and imported into the VR environment. This reduces major challenges to selecting a software platform to develop the VR environment, graphics processing unit (GPU), head mounted display (and its software), hand-gesture sensor (and its software), computer, and chair assembly. When full-body sensing hardware and software become available at a reasonable price, we will explore incorporating them into OVAL.

For building the VR environment, perhaps the most important choice is selecting the software platform. Currently, game engines (software platforms for generating 2D and immersive 3D environments) are the best option. To select an appropriate game engine, we established five criteria:

1. Minimally priced or free during development of VR environment

2. Compatible with multiple platforms, including desktop, mobile, and Web

3. Compatible with multiple headsets, such as Oculus Rift, HTC Vive, Google Cardboard, Gear VR, and Microsoft HoloLens

4. Active and robust developer community

5. Extensive online documentation.

Our solution

Our criteria limited our choices to two main video game engines: Unity3D and Unreal. Although, since our initial choice, a number of major companies (Google, Amazon, etc.) have developed game engines, we still champion Unity3D because of its low cost and robust developer community, which attests to its capability and compatibility. Part of Unity3D's appeal is one of our main criteria: strong cross-platform compatible. Currently, it is compatible with 21 different operating systems, including Apple, Windows, Linux, and a variety of mobile devices. Furthermore, Unity3D provides strong online documentation, including tutorials.

For software sub-systems, choices had their complexities, and we will discuss them more fully in the presentation. They included Photon Unity Networking for networking the headsets; Oculus runtime software to support a camera with two "eyes"; and the LeapMotion SDK for hand-tracked interactions. These choices allow OVAL to preserve embodied interaction with hand and upper body tracking (leaning in produces a closer look at an object, and hand movements control features such as scaling and rotating objects) (Shapiro, 2014).

Selecting hardware was less complicated. Available headsets are surprisingly limited. Our only real choice was Oculus Rift. Another headset, the HTC Vive, is set for release soon. Most headsets are developed and sold as part of a complete VR system.

For the chair-assembly, a unique on-campus resource simplified our choice. The University of Oklahoma houses a high-powered physics fabrication lab. We worked with them to develop a custom railed-chair assembly (ergonomically designed for a 360° range of motion). This railed-chair allows the computer to reside under the chair and out of the way. For a robust virtual environment, the computer contains a GeForce GTX 980 graphics card. It delivers a 75-frames/second refresh-rate (the human eye generally resolves 25 frames/second), insuring an instantaneous visual experience when manipulating 3D objects or when turning one's head.

Finally, by integrating networking software into OVAL, a shared VR experience can occur across a range of clients. All changes made on a master workstation—including scale, rotation, lighting, and background imagery—are immediately transmitted to all co-participants, regardless of their physical location. In a classroom environment, for example, this means that students automatically see what the teacher sees. But this also allows OVAL to become a worldwide network. To facilitate such a network, all 3D models are uploaded via a public Dropbox, which immediately syncs with all OVAL clients. This means that all uploaded 3D asset are available to all OVAL clients. For a shared VR experience, each client only needs a short set of instructions concerning file names and how to manipulate them during a session.

Research and teaching

In our presentation, we will also discuss ongoing uses of OVAL at the University of Oklahoma and explore their implications. Despite its recent completion, OVAL has already had extensive use. Undergraduate biology students have analyzed the atomic structure of hemoglobin and oxyhemoglobin. Architecture faculty has analyzed student projects for unseen flaws pertaining to safety and accessibility of interior spaces. The Sam Noble Museum of Natural History has uploaded their recently discovered *Aquilops Americanus* skull into the OVAL system for curators and researchers. Art History faculty has begun analyzing sculpturally significant 3D scans for preserving what was once ephemeral art. A budding partnership with the Medical Imaging Facility has demonstrated how CT-to-OVAL workflows facilitate mammographic research. Finally, Bill Endres has begun to develop guided, immersive tours of the St Chad Gospels, an 8th-century illuminated manuscript.

Conclusion

The rapid production of 3D models makes having VR systems available for their viewing a pressing concern. 3D models of massive structures, such as the large Buddhas of Bamiyan, highlight the limitations of interacting through a computer screen. OVAL provides one cost-efficient solution. In our next phase, we plan to add collaborators and make OVAL available. We are also interested in hosting 3D assets in an archive-quality database. However, the most effective and efficient means of doing these has yet to be determined. We are looking forward to presenting at DH 2016 and conversing about possibilities for OVAL and the wide-ranging opportunities for research and teaching through VR.

Bibliography

Dylla, K., et al. (2010). Rome Reborn 2.0: A Case Study of Virtual City Reconstruction Using Procedural Modeling Techniques. Frischer, B., Crawford, J. and Koller, D. (eds), *Making History Interactive: 37*th *Proceedings of the CAA Conference*, Williamsburg, VA, March 2009. Oxford: Archaeopress, pp. 62-66.

Grün, A., Remondino, F. and Zhang, L. (2004). Photogrametric Reconsctruction of the Great Buddha of Bamiyan, Afghanistan. *The Photogrammetric Record*, **19**(107): 177-99.

Hermon, S. and Kalisperis, L. (2011). Between the Real and the Virtual: 3D Visualization in the Cultural Heritage Domain – Expectations and Prospects. *Virtual Archaeology Review*, **2**(4): 59-63.

Shapiro, L. (ed) (2014). *The Routledge Handbook of Embodied Cognition*. New York: Routledge.

Outliers or Key Profiles? Understanding Distance Measures for Authorship Attribution

Stefan Evert
stefan.evert@fau.de
Universität Erlangen-Nürnberg, Germany

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

Thomas Proisl
thomas.proisl@fau.de
Universität Erlangen-Nürnberg, Germany

Thorsten Vitt
thorsten.vitt@uni-wuerzburg.de
University of Würzburg, Germany

Christof Schöch
christof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Isabella Reger
isabella.reger@uni-wuerzburg.de
University of Würzburg, Germany

The state of the art

Burrows' Delta is one of the most successful algorithms in computational stylistics (Burrows 2002). A series of studies have proven its usefulness (e.g. Hoover 2004, Rybicki & Eder 2011). There are two essential steps in Burrows' Delta. The first is to standardize the relative frequencies of words in a document-term-matrix through a z-score transformation. In the second step, the distances between all texts are calculated. For each word, the difference between the z-score of the word in one and the other text are calculated. The absolute values of the differences are added for all words taken into account. The usual interpretation is that the smaller the sum, the more similar two texts are stylistically, and the more likely it is that they have been written by the same author.

Despite the fact that Burrows' Delta is as simple as it is useful, there is still a lack of a good explanation why the algorithm works so well. Argamon (2002) has shown that the second step in Burrows' Delta is equivalent to taking the Manhattan distance between two points in a multi-dimensional space. He suggests, among other things, using the Euclidean distance instead. An empirical test of his proposals has shown, however, that none of them lead to an improvement in performance (Jannidis et al. 2015).

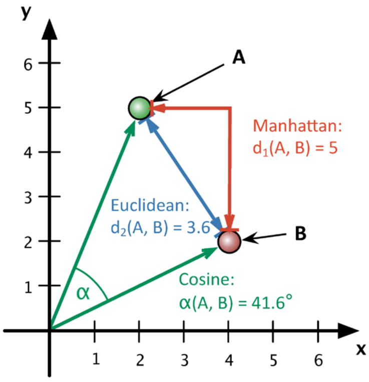


Figure 1: Illustration of the distance between two texts made up of just two words

Smith and Aldrige (2011) have suggested to use the cosine of the angle between the document vectors for the second step, as is customary in Information Retrieval (Baeza-Yates & Ribeiro-Neto 1999:27). The Cosine variant of Delta (Delta *Cos*) outperforms Burrows' Delta (Delta *Bur*) in many different settings and has the advantage of not showing the drop in performance typical ofother Delta variants when large numbers of MFW are used (Jannidis et al. 2015). The question now is why Delta *Cos* is so much better than Delta *Bur* and other variants, that is, in what way Delta *Cos* captures the authorship signal more clearly than other variants of Delta.

Of decisive importance for our further analyses was the insight that using the Cosine Distance is equivalent to a vector normalization in the sense that (in contrast to Manhattan and Euclidean Distance) the length of the vector does not play a role for the calculation of the distance (see figure 1). Previous experiments have shown that an explicit, additional vector normalization also substantially improves performance of the other Delta measures (Evert et al. 2015).

Hypotheses

Having discovered that impact of the normalization effect, we have developed two empirically testable hypotheses:

- (H1) Performance differences are caused by single extreme values, so-called outliers. These are particularly large positive or negative *z-scores* specific to single texts rather than all texts of a single author. As the Euclidean distance should be more sensitive to single extreme values than the Manhattan distance, this hypothesis would explain the comparatively bad performance of Argamon's "Quadratic Delta" Delta *Q*. The positive effect of vector normalization originates from the reduction of outlier amplitudes ("outlier hypothesis").
- (H2) The author specific "style profile" manifests itself more in the qualitative combination of word preferences, i.e. in the pattern of over- and under utilization of vocabulary, rather than in the actual amplitude of *z-scores*. A text distance measure is particularly successful in authorship attribution if emphasizing structural differences of author style profiles without being too much influenced by actual amplitudes ("key-profile hypothesis"). This hypothesis explains directly why vector normalization results in such impressive improvements: it standardizes the amplitudes of author profiles in different texts.

New insights

Corpora

For the experiments in this paper, we use three similarly composed corpora in German, English and French. Each corpus contains 25 different authors with 3 novels each, thus 75 texts in total. The corpora have been described in Jannidis et al. (2015). Due to space issues, the following section will only present our observations on the German corpus. The results for the corpora in both other languages show only small deviations and also support our findings.

Experiments

To further investigate the role of outliers and thus the plausibility of H1, we complement Delta *Bur* and Delta *Q* with additional variants based on the general Minkowski distance (for $p \geq 1$):

$$\Delta_p = \left(\sum_{i=1}^{m} |z_i(D_1) - z_i(D_2)|^{p} \right)^{1/p}$$

We generally name these distance measures L *p*-Delta. The specific case $p = 1$ equals the Manhattan distance (L *1*-Delta = Delta *Bur*), $p = 2$ the Euclidean distance (L *2*-Delta = Delta *Q*). The higher the value for *p*, the larger the influence of single outliers on L *p*-Delta.

Fig. 2 compares four different L *p* distance measures (for p=1, √2, 2, 4) with Delta *Cos*. The method of comparison is the same as in Evert et al. (2015): 75 text are automatically clustered in 25 groups according to Delta distances; clustering quality is estimated with the adjusted rand index (ARI). An ARI of 100% signifies perfect author recognition whereas a value of 0% shows that the clustering is entirely random. The performance of L *p* Delta obviously decreases with increasing *p*. Additionally, the robustness of the measures also decreases with an increasing number of MWF used. As already reported in Jannidis et al. (2015) and Evert et al. (2015), Delta$_{\text{Bur}}$ (L_1) consistently outperforms Argamon's Delta *Q* (L 2). Especially if many features, i.e. a large number of MFW is considered, high p values result in low performance. Delta *Cos* is more robust than other variants and achieves almost perfect attribution success (ARI > 90%) over a wide range of the MFW.

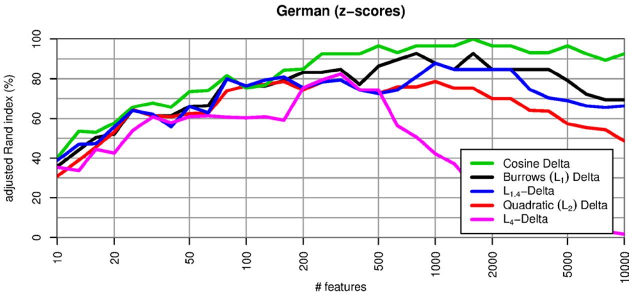


Figure 2: Clustering quality of different Delta measures as a function of the number of the MFW considered

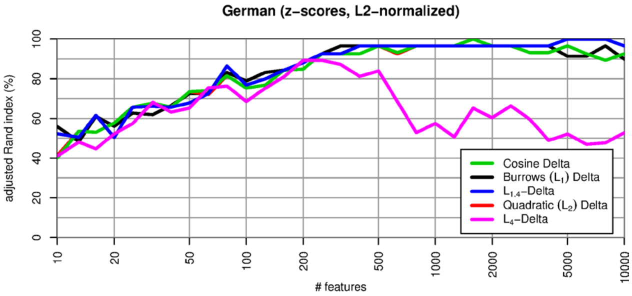


Figure 3: Cluster quality of various Delta measures with length-normalized vectors

Normalizing the feature vectors to length 1 improves the quality of all Delta measures significantly (fig. 3). In this case, Argamon's Delta *Q* is identical to Delta *Cos*: the red line is completely covered by the green one. The other Delta measures (Delta *Bur*, L *1.4*-Delta) now reach about the same quality as Delta *Cos*. Only L *4* Delta, which is especially prone to outliers, falls short considerably. These results seem to support H1.

A different approach to limit the influence of outliers is to truncate extreme *z-scores*. To do so, we set all $|z| > 2$ to +2 or −2, depending on the original *z-scores*'s sign. Fig. 4 shows the effects of various normalizations on the distribution of the feature values. Vector length normalization (lower left) produces only slight changes and practically does not reduce the number of outliers at all. Pruning large *z-score* values only affects words with above-average frequencies (upper right).

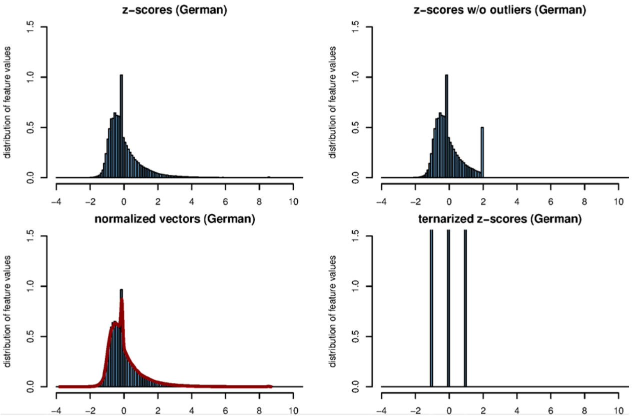


Figure 4: Distributions of feature vectors for all 75 texts, using vectors of 5000 most frequent words. The table shows the distribution of the original *z-scores* (upper left), the distribution after length-normalizing the vectors (lower left), the distribution after clamping outliers with $|z| > 2$ (upper right) and a ternary quantization to the values −1, 0 and +1 (lower right). The red curve in the lower left graph shows the *z-scores* before normalization; the direct comparison shows the normalization has only minimal effect and almost does not reduce outliers. The thresholds for the ternary quantization, $z < -0.43$ (−1), $-0.43 \leq z \leq 0.43$ (0) and $z > 0.43$ (+1), have been selected such that in an ideal normal distribution, a third of all feature values would fall into each of the classes −1, 0, and +1.

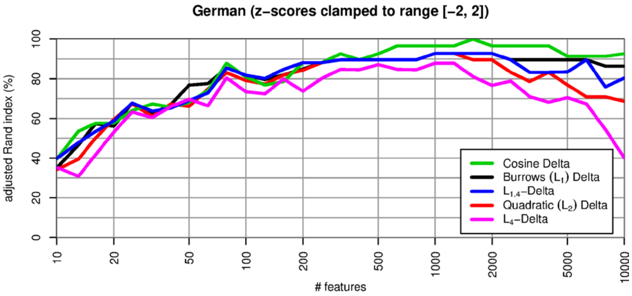


Figure 5: Cluster quality after clamping outliers, i.e. feature values with $|z| > 2$ have been replaced with the fixed values −2 or +2, depending on *z-score*'s sign

As Fig. 5 shows, this manipulation improves the performance of all L *p* Deltas considerably. However, its positive effect is noticeably smaller than that of vector normalization.

With these differing effects of the normalizations on outlier distributions and Delta results, H1 cannot be upheld. H2 is supported by the good results of vector length normalization. However, on its own, it cannot explain why clamping outliers leads to a considerable improvement as well. To examine this hypothesis further, we created pure "key profile" vectors that only discriminate between word frequencies that are above average (+1), unremarkable (0), and below average (−1; cf. Fig. 4, lower right).

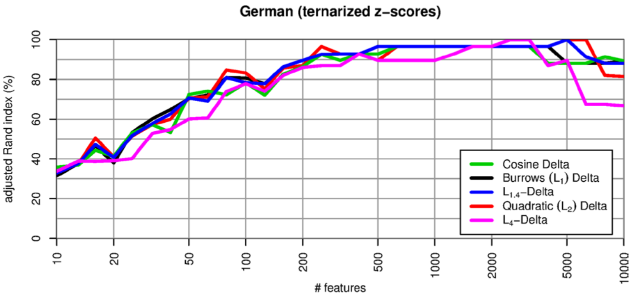


Figure 6: Cluster quality with ternary quantization of the vectors in frequencies that are above average (+1, $z > 0.43$), unremarkable (0, $-0.43 \leq z \leq 0.43$), and below average ($z < -0.43$)

Fig. 6 shows that these key profile vectors perform remarkably well, almost on par with vector normalization. Even the especially outlier-prone L *4* Delta reaches a quite robust clustering quality of more than 90%. We interpret this observation as giving considerable support to hypothesis H2.

Discussion and perspectives

H1, the outlier hypothesis, has been disproven as the vector normalisation hardly reduces the number of extreme values and the quality of all L *p* measures is still considerably improved. On the other hand, H2, the key profile hypothesis, has been confirmed. The ternary quantification of the vectors shows clearly that it is not the extent of deviation resp. the size of the amplitude, but the profile of deviation across the MFW which is important. Remarkably, the measures behave differently if more than 2000 MFW are used. Almost all variant show a decline for a very large number of features, but they differ in when this decline starts. We suppose that the vocabulary in those parts is less specific for an author than for topics and content. Clarifying such questions will require further experiments.

Bibliography

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, **23**(2): 131–47 doi:10.1093/llc/fqn003. http://llc.oxfordjournals.org/content/23/2/131.abstract.

Baeza-Yates, R. and Ribeiro Neto, B. (1999). *Baeza-Yates, Ricardo; Ribeiro Neto, Berthier (1999): Modern Information Retrieval. Harlow.* Harlow.

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, **17**(3): 267–87 doi:10.1093/llc/17.3.267. http://llc.oxfordjournals.org/content/17/3/267.abstract.

Eder, M. and Rybicki, J. (2011). Deeper Delta across genres and languages: do we really need the most frequent words?. *Literary and Linguistic Computing*, **26**(3): 315–21 doi:10.1093/llc/fqr031. http://llc.oxfordjournals.org/content/early/2011/07/14/llc.fqr031.abstract .

Evert, S., Proisl, T., Pielström, S., Schöch, C. and Vitt, T. (2015). Towards a better understanding of Burrows's Delta in literary authorship attribution. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver CO.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75 doi:10.1093/llc/19.4.453. http://llc.oxfordjournals.org/content/19/4/453.abstract.

Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015). Improving Burrows' Delta – An empirical evaluation of text distance measures. *Digital Humanities 2015 Conference Abstracts*. Sydney: ADHO http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html.

Smith, P. W. H. and Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. *Journal of Quantitative Linguistics*, **18**(1): 63–88 doi:10.1080/09296174.2011.533591. http://www.tandfonline.com/doi/abs/10.1080/09296174.2011.533591.

Die digitale Modellierung experimenteller Druckgrafiken des 15. Jahrhunderts. Anforderungen und Chancen computerbasierter Dokumentationsverfahren

Peter R. Fornaro
peter.fornaro@unibas.ch
University of Basel, Switzerland

Lothar Schmitt
lschmitt.basel@gmail.com
University of Basel, Switzerland

Andrea Bianco
andrea.bianco@unibas.ch
University of Basel, Switzerland

Heidrun Feldmann
heidrun.feldmann@unibas.ch
University of Basel, Switzerland

Das Forschungsprojekt “Digitale Materialität” des Schweizerischen Nationalfonds ist eine Kooperation des Digital Humanities Labs und des Kunsthistorischen Seminars der Universität Basel. Ziel dieses Projektes ist es, Verfahren zur digitalen Reproduktion von Kunstwerken zu entwickeln sowie bereits bestehende Methoden zu optimieren, um sie in der kunsthistorischen Forschung besser einsetzen zu können.

Ein Themenbereich, in dem diese Vorgehensweise erprobt wird, sind frühe experimentelle Druckgrafiken, die im 15. Jahrhundert entstanden sind. Sie sind Teil der Medienrevolution, die der Erfindung des Buchdrucks folgt. Anders als konventionelle Graphiken nutzen die experimentellen Drucktechniken Farb- und Reliefeffekte, um attraktiver und wertvoller zu wirken. Beides sind Oberflächeneigenschaften die sich nur schwer fotografisch aufnehmen lassen.

Zu diesen graphischen Experimenten gehören die sogenannten Teigdrucke (Mabbott, 1932; Bowman, 1985; Bertalan, 1993; Scheld, 2009). Sie bestehen aus mehreren übereinanderliegenden Schichten. Ein Papieruntergrund ist mit einer verformbaren Substanz bedeckt. Sie wurde mit einer dünnen Metallfolie überzogen, um den Eindruck zu erwecken, als sei Gold verwendet worden. In diesen mehrschichtigen "Teig" wurde eine Metallplatte gepresst, in die man zuvor das abzubildende Motiv graviert hatte. Die Oberfläche der Platte wurde für den Druckprozess mit schwarzer Farbe bedeckt. Dieses Schwarz übertrug sich beim Drucken auf den golden schimmernden Untergrund.

Solche Drucke sind sehr selten, ausgesprochen fragil und deshalb meist in schlechtem Zustand erhalten. Ihre Oberfläche ist an vielen Stellen abgebröckelt. Auch der metallische Glanz ist nur noch an wenigen Stellen sichtbar. Darum fällt es schwer, sich eine Vorstellung vom einst so filigranen, kostbar schimmernden Relief zu machen. Aufgrund der Oberflächenkomplexität wirken Teigdrucke auf konventionellen Fotografien unansehnlich, obwohl sie in Wirklichkeit einiges von ihrer faszinierenden Materialwirkung bewahrt haben.

Um diese interessanten Objekte besser reproduzieren zu können, verbindet das Projekt “Digitale Materialität” Komponenten aus der Fotografie und der Computergrafik, welche es ermöglichen, die materiellen Eigenschaften der Originale so abzubilden, dass sie mit der beobachtbaren Realität in einem überprüfbaren Zusammenhang stehen. Um das ursprüngliche Aussehen in der digitalen Domäne zu representieren, wird für die fotografische Aufnahme das sogenannte Reflection Transformation Imaging (RTI) eingesetzt (Malzbender, Gelb, Wolters, 2001; Mudge et al., 2008; MacDonald, 2015). Mit diesem Verfahren wird aus mehreren statischen Aufnahmen ein mathematisches Relexionsmodell gerechnet, welches erlaubt die Reflexionseigenschaften der Oberfläche eines Objekts interaktiv darzustellen. So wird es möglich, die Reflexe unter sich verändernder Beleuchtung verlässlich zu simulieren. Die grundsätzliche Methodik von RTI besteht im systematischen Aufnehmen von Fotografien unter unterschiedlichen Lichteinfallswinkeln. Dieses Aufnahme-Set dient als Datenbasis, um eine mathematische Funktion so zu parametrisieren, dass sie die physikalischen Gegebenheiten für jeden Ort (jedes Pixel) möglichst exakt wiedergeben kann. Das auf diese Weise generierte Modell kann anschliessend am Computer betrachtet werden, wobei sich das Aussehen des Objekts unter wechselnden Lichtsituationen interaktiv simulieren lässt.

Die erreichten Ergebnisse zeigen, dass RTI die Oberfläche, ihre Reflexionseigenschaften und den aktuellen Erhaltungszustand von Teigdrucken aussagekräftiger dokumentieren und darstellen kann, als konventionelle Fotografien. Insbesondere das feine dreidimensionale Relief der Teigdrucke lässt sich an RTI-Modellen besser nachvollziehen.

Im Rahmen des Basler Projekts wird das Verfahren aber nicht nur angewendet, sondern auch weiterentwickelt. Das konventionelle RTI Verfahren stösst bei Objekten mit kombinierten glänzenden und matten Oberflächen an seine Grenzen. Dies liegt am einfachen mathematischen Modell – im Normalfall eine einfache Parabel – sowie an der Tatsache, dass das selbe Modell für das ganze Bild verwendet wird, also nicht auf einen ggf. vorhandenen Materialmix eingegangen wird. Durch das Überlagen eines aus der Computergrafik stammenden Glanzmodells – z.B. Phong (Phong, 1975) oder Ward (Ward and Glencross, 2009) – kann dieses Manko behoben werden, um so dem Reflexionsverhalten unterschiedlicher Materialien Rechnung zu tragen. Die Weiterentwicklungen werden laufend von kunsthistorischer Seite begleitet, um zu bewerten, wie die verbesserten Modelle im Vergleich zu den bislang verwendeten Verfahren als neues Analyseinstrument für geisteswissenschaftliche Forschungen genutzt werden können.

Um schliesslich die im Projekt generierten digitalen Objekte der Forschung zur Verfügung zu stellen, werden sie in ein Virtual Research Environment (Rosenthaler and Subotic 2012, Carusi and Reimer, 2010) integriert. Zur Wiedergabe der Modelle setzen wir auf WebGL (https://www.khronos.org/webgl/). Die Verbreitung der Daten im Rahmen von Open Access Anforderungen wird durch eine standardisierte Schnittstelle (JSON, JSON-LD) realisiert.

Am Beispiel der Teigdrucke zeigt sich im Basler Projekt, wie Entwicklungen der Digital Humanities Voraussetzung für neue Forschungsfragen in den Geisteswissenschaften schaffen können. An die Stelle des erprobten Dokumentations- und Vermittlungsmediums der konventionellen Fotografie tritt mit RTI ein innovatives Verfahren. Es wird jedoch nicht nur zweckmässig eingesetzt, sondern muss anwendungsbezogen optimiert werden, um den hohen Ansprüchen kunsthistorischer Dokumentationspraxis gerecht zu werden. Gefordert sind nämlich nicht etwa multimediale sondern gänzlich neuartige Forschungsdaten, die andernfalls für wissenschaftliche Diskurse nicht verfügbar wären.

Hinzu kommt die Aufbereitung des Verfahrens für den Gebrauch in einer neuen, virtuellen Forschungsumgebung, die ihrerseits traditionelle Formen der wissenschaftlichen Vernetzung ergänzt. Dies geschieht unter anderem durch den Einsatz semantischer Technologien, die nur im digitalen Raum ihr Potential entfalten können. Die virtuelle Forschungsumgebung wird schliesslich in Kombination mit webfähigen Applikationen und Open Access-Schnittstellen zum Angelpunkt für eine dauerhafte Sicherung der digitalen Daten, die zugleich nutzbar und archivierbar gemacht werden.

Bibliographie

Bowman, C. L. (1985). Pasteprints. A New Hypothesis About Their Production. *Print Quarterly* **2**: 4-11.

Bertalan, S. M. (1993). Medieval pasteprints in the National Gallery of Art. In Ross M. Merrill u. a. (eds.), *Conservation research*. Hanover u. a.: University Press of New England, pp. 30-61.

Carusi, A. and Reimer, T. (2010). Virtual Research Environment Collaborative Landscape Study *JISC*. http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf [letzter Zugriff 26. Februar 2016].

Mabbott, T. O. (1932). Pasteprints and Sealprints. *Metropolitan Museum Studies* **4**: 55 -75.

MacDonald, L. (2015). *Realistic visualisation of cultural heritage objects*. London: University College 2015.

Malzbender, T., Gelb, D. and Wolters, H. (2001). Polynomial Texture Maps. *SIGGRAPH 01, Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. New York: ACM Press 2001, pp. 519-28.

Mudge, M. et al. (2008). Image-Based Empirical Information Acquisition, Scientific Reliability, and Long-Term Digital Preservation for the Natural Sciences and Cultural Heritage. *EUROGRAPHICS*. http://culturalheritageimaging.org/What_We_Do/Publications/eurographics2008/eurographics_2008_tutorial_notes.pdf [letzter Zugriff 26. Februar 2016].

Phong, B. T. (1975). Illumination for Computer Generated Pictures. *Commun*. ACM **18**(6): 311-17.

Rosenthaler, L. and Subotic, I. (2012). SALSAH 2.0 – Augmenting a virtual research environment with source-centric analysis methods. SNF-Proposal no. 137929: Bern.

Scheld, A. and Damm, R. (2009). Flock Prints and Paste Prints. A Technological Approach. *Peter Parshall* (ed.): The Woodcut in Fifteenth-Century Europe. New Haven: Yale University Press 2009, pp. 316-36.

Ward, G. and Glencross, M. (2009). *A case study evaluation: perceptually accurate textured surface models*. Manchester: ACM Press: 109.

REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions

Francesca Frontini
francesca.frontini@ilc.cnr.it
Istituto di Linguistica Computazionale "A. Zampolli" - CNR Pisa

Carmen Brando
carmen.brando@ign.fr
Institut National de l'Information Géographique et Forestière - Paris

Jean-Gabriel Ganascia
jean-gabriel.ganascia@lip6.fr
Labex OBVIL - LIP6 (Laboratoire d'Informatique de Paris 6)
Université Pierre et Marie Curie and CNRS

Introduction

As Susan Schreibman (2014) points out, a digital edition, as opposed to a printed one, is never really complete as several layers of annotation may always be added to represent and enrich the original content. TEI (Burnard, 2014) allows for several types of information - textual, linguistic and semantic - to be layered and made explicit and retrievable by a machine. Such is the case for instance with what is commonly known as semantic tagging.

In this paper, we focus on Named Entities (NE), in particular names of Persons and Geographical Places. Adding

NE mentions is supported by TEI with appropriate tags (such as <persName> and <placeName>), whose addition in a digital critical edition has somewhat the same function that indexes of places and persons have in a printed one. As mentions may be ambiguous (same string for different people, same place with different names,....) some referencing and disambiguating identifiers are required. But digital editions allow for much more than simple internal referencing. By pointing to external sources, structured information contained in the form of linked data in the semantic web becomes available to scholarly research.

In this work we present REDEN ONLINE, a system that enables scholars to automatically add external references to annotations of persons and places. The system is a web interface taking TEI as input, where mentions are already marked up, and automatically disambiguates and links such entities to an appropriate linked data set using a graph based algorithm for disambiguation. Moreover, our system provides data aggregation and visualization facilities by using the information found in the reference sources.

Previous work and general context

Semantic tagging is a hot topic in the digital humanities. Tools for semantic enrichment are, such as *Pundit* (Grassi et al., 2012, 2013), already available and allow for the interactive and intuitive annotation of portions of text. Automatic Named Entity Recognition and Linking techniques may be implemented to detect mentions and to suggest links to external knowledge bases.

Input formats to such systems may vary from plain text to html, but ideally a tool should process available standard formats, such as TEI-XML for text and RDF/OWL for information. Using linked data sources for disambiguation and enrichment is thus strongly recommended. By doing this, external sources of structured and regularly updated information can be made available to the scholar without having to be directly incorporated into the inline annotation, that can be left as simple as possible. This in turn allows for several customizable views, as linked data sources may be queried with the SPARQL query language to retrieve only the amount of external information that is necessary for a given task.

The treatment of spatial and temporal information is a typical task for which this approach is particularly effective; the availability of geographical databases and the complexity of the information are best accessed by pointing from within the digital edition to an external link. But also other types of semantic information seem to be particularly apt for connection to rich linked databases. So for instance bibliometric sources can be used to enrich texts with additional information on authors.

Typical targets for references are DBpedia and Geonames, that, for their genericity and connection to other sources, are at the heart of the linked data cloud. But they may be supplemented by more domain specific sources of information. For instance, Pleiades provides geo-historical information for ancient places.

Our project

REDEN ONLINE is set against the background of work carried out at LABEX OBVIL in Paris, where quality digital editions for French literary texts and criticism are produced and used in research and higher education. Recently a series of projects were carried out to semi-automatically annotate and reference places, organizations and authors. Gold standards were also produced, in close contact with researchers in French literature, so as to establish guidelines of annotation that best suit their ongoing research.

The general purpose is to provide tools for both:

- augmented close reading, to enable researchers to access more information on a specific text portion
- distant reading and data aggregation, so as to be able to detect trends in large portions of texts (Moretti, 2007)

OBVIL literary scholars are interested in plotting the distribution of the mentions of given authors over time in French literary criticism, in order to study the appreciation of Molière over the centuries, or in producing charts representing the distributions of professions in authors mentioned in given periods, to trace the influence of scientists and their ideas on art and literature in the age of positivism (Riguet, 2015). Other visualizations captured the emerging influence of foreign countries in the French literary panorama over time by combining the date of the publication of the essays with the detected toponyms.

NLP technologies are used to facilitate various aspects of the semantic enrichment of TEI editions, in an annotation echosystem where texts are first processed and then manually checked. The detection of mentions of places, authors (and also organizations) was tackled by using a Named Entity Recognizer and Classifier (UNERD, Mosallam et al., 2014).

Once the entities are correctly detected and classified, external references need to be added to disambiguate mentions and to connect them to additional information. To this purpose we developed REDEN[1], a Named Entity Linker that uses a graph-based algorithm and linked data sets to identify the correct referent for each mention (Brando et al., 2015, Frontini et al., 2015a, Frontini et al., 2015b for the technical details).

REDEN's input consists of a TEI text with detected mentions and several parameters specifying among others the class of entities to be detected, the reference base to use and a set of pre-compiled indexes. REDEN is applied for each class of entities separately, and works at best when several mentions are disambiguated at the same time. It retrieves all candidate referents for each mention of a context (say a paragraph) and then all the available information from the semantic web. It builds a sub-graph of all candidates and chooses the correct referents for each mention with the help of the formal relations between them. From Figure 1 you can get an intuition of how REDEN works.

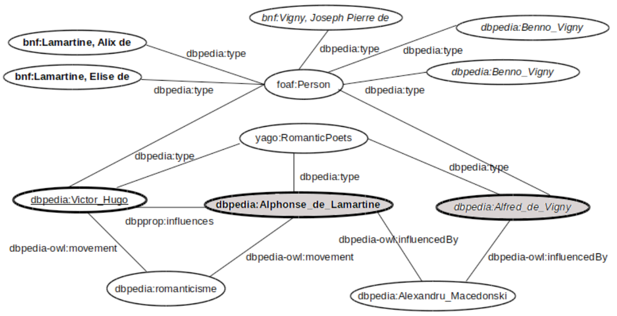


Figure 1 The graph based algorithm disambiguates between different possible referents for the mentions of "Victor Hugo" (unambiguous in this example), "Lamartine" and "Vigny" (both having several candidate referents) based on information found in DBpedia. Correct referents (in grey) are chosen based on how well connected they are within the context Here the crucial node is clearly that of yago:RomanticPoets.

So far our efforts have concentrated on the production of a text annotation and referencing pipeline for the production of such enriched TEIs with annotated and referenced mentions. Their exploitation for data aggregation and visualization was carried out offline and with ad hoc processing tools. With REDEN ONLINE we now want to make linking technology available online while at the same time providing users with some generic visualization of the results.

In what follows, we present the REDEN ONLINE interface with some screenshots from an example where two texts of the Labex OBVIL[2] digital library have been automatically linked to external sources, namely:

- *L'Hérésiarque et cie*, a collection of short stories by Guillaume Apollinaire, published in 1910 - place mentions linked to DBpedia entries.

- Réflexions sur la littérature a series of essays on French literary criticism by Albert Thibaudet, published in 1936 - author's mentions linked to entries in the linked data base of the Bibliothèque Nationale de France (BnF).

The user (Figure 2) loads a TEI text with annotated <placeName> or <personName> tags, chooses which class of entities to process (places or nouns) and the system runs the disambiguation and linking algorithm against the given linked data base - here French DBpedia and/or BnF. Then external information is extracted from the source and used for generating a particular view of the text. The result is a summing up of the disambiguated locations (some place names may be non resolvable because they are absent from the linked data base) and a visualization.

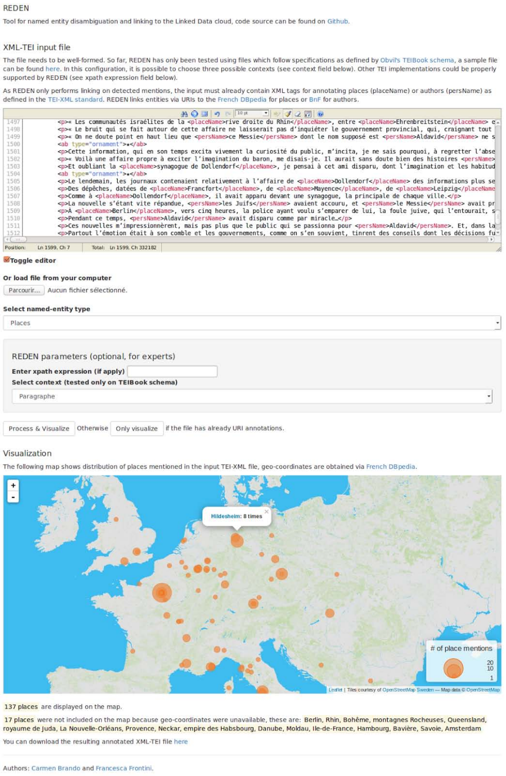
For locations the visualization consists in an interactive map that also takes frequency of mention into account. Coordinates are retrieved from DBpedia when available and the map can be zoomed in, up to the level of streets (see Figure 3 where some places in Paris have been identified in the text by Apollinaire), when relevant.

For persons (see Figure 4), portraits of authors are automatically downloaded and visualized.

Conclusion

The conference presentation will demonstrate REDEN ONLINE, a web based tool that enables researchers to connect place names and person names in their texts to existing linked data sources. The underlying technology will also be explained, in particular its use of standard formats, such as TEI and RDF for the linking algorithm, and GeoJSON for the creation of the map. We will also argue in favour of our economicity approach, namely the choice of not embedding semantic information in the TEI, which enables the use of different databases and the production of ad hoc "views" of the document.

It is well known that aggregation and visualizations techniques may "assist the critic in the unfolding of interpretive possibilities" (Ramsay, 2008) when analysing texts. This tool has been particularly designed for the study of literature and literary criticism; in the presentation examples of use will be given using ongoing research on Apollinaire, highlighting how the visual representation of the itineraries contained in the stories may be considered as a form of novel "digital reading" of the text.



REDEN

Tool for named entity disambiguation and linking to the Linked Data cloud, code source can be found on Github.

XML-TEI input file

The file needs to be well-formed. So far, REDEN has only been tested using files which follow specifications as defined by Obvil's TEIBook schema, a sample file can be found here. In this configuration, it is possible to choose three possible contexts (see context field below). Other TEI implementations could be properly supported by REDEN (see xpath expression field below).

As REDEN only performs linking on detected mentions, the input must already contain XML tags for annotating places (placeName) or authors (persName) as defined in the TEI-XML standard. REDEN links entities via URIs to the French DBpedia for places or BnF for authors.

```
        <p>« Les communautés israélites de la <placeName>rive droite du Rhin</placeName>, entre <placeName>Ehrenbreitstein</placeName> e
        <p>« Le bruit qui se fait autour de cette affaire ne laisserait pas d'inquiéter le gouvernement provincial, qui, craignant tout
        <p>« On ne doute point en haut lieu que <persName>ce Messie</persName> dont le nom supposé est <persName>Aldavid</persName> ne s
        <ab type="ornament">*</ab>
        <p>Cette information, qui en son temps excita vivement la curiosité du public, m'incita, je ne sais pourquoi, à regretter l'abse
        <p>« Voilà une affaire propre à exciter l'imagination du baron, me disais-je. Il aurait sans doute bien des histoires <persName>
        <p>Et oubliant la <placeName>synagogue de Dollendorf</placeName>, je pensai à cet ami disparu, dont l'imagination et les habitud
        <ab type="ornament">*</ab>
        <p>Le lendemain, les journaux contenaient relativement à l'affaire de <placeName>Dollendorf</placeName> des informations plus se
        <p>Des dépêches, datées de <placeName>Francfort</placeName>, de <placeName>Mayence</placeName>, de <placeName>Leipzig</placeName
        <p>Comme à <placeName>Dollendorf</placeName>, il avait apparu devant une synagogue, la principale de chaque ville.</p>
        <p>La nouvelle s'étant vite répandue, <persName>les Juifs</persName> avaient accouru, et <persName>le Messie</persName> avait pr
        <p>À <placeName>Berlin</placeName>, vers cinq heures, la police ayant voulu s'emparer de lui, la foule juive, qui l'entourait, s
        <p>Pendant ce temps, <persName>Aldavid</persName> avait disparu comme par miracle…</p>
        <p>Ces nouvelles m'impressionnèrent, mais pas plus que le public qui se passionna pour <persName>Aldavid</persName>. Et, dans la
        <p>Partout l'émotion était à son comble et les gouvernements, comme on s'en souvient, tinrent des conseils dont les décisions fu
```

Position: Ln 1599, Ch 7 Total: Ln 1599, Ch 332182

☑ Toggle editor

Or load file from your computer

Parcourir... Aucun fichier sélectionné.

Select named-entity type: Places

REDEN parameters (optional, for experts)

Enter xpath expression (if apply)

Select context (tested only on TEIBook schema): Paragraphe

Process & Visualize Otherwise Only visualize if the file has already URI annotations.

Visualization

The following map shows distribution of places mentioned in the input TEI-XML file, geo-coordinates are obtained via French DBpedia.

10 pt
+
-
Hildesheim: 8 times
# of place mentions
20
10
1
Leaflet | Tiles courtesy of OpenStreetMap Sweden — Map data © OpenStreetMap

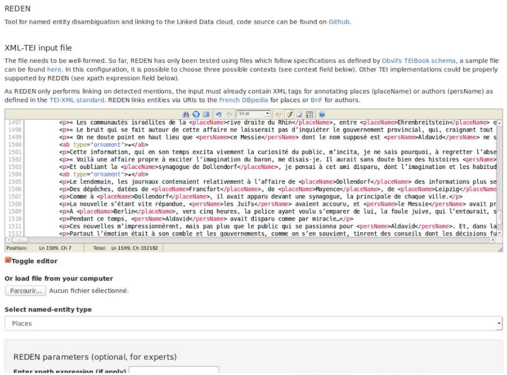
137 places are displayed on the map.

17 places were not included on the map because geo-coordinates were unavailable, these are: Berlin, Rhin, Bohême, montagnes Rocheuses, Queensland, royaume de Juda, La Nouvelle-Orléans, Provence, Neckar, empire des Habsbourg, Danube, Moldau, Ile-de-France, Hambourg, Bavière, Savoie, Amsterdam

You can download the resulting annotated XML-TEI file here

Authors: Carmen Brando and Francesca Frontini.

Figure 2 The REDEN ONLINE interface, with a sample text from Apollinaire. Place names results are visible as a map visualization.



REDEN

Tool for named entity disambiguation and linking to the Linked Data cloud, code source can be found on Github.

XML-TEI input file

The file needs to be well-formed. So far, REDEN has only been tested using files which follow specifications as defined by Obvil's TEIBook schema, a sample file can be found here. In this configuration, it is possible to choose three possible contexts (see context field below). Other TEI implementations could be properly supported by REDEN (see xpath expression field below).

As REDEN only performs linking on detected mentions, the input must already contain XML tags for annotating places (placeName) or authors (persName) as defined in the TEI-XML standard. REDEN links entities via URIs to the French DBpedia for places or BnF for authors.

```
        <p>« Les communautés israélites de la <placeName>rive droite du Rhin</placeName>, entre <placeName>Ehrenbreitstein</placeName> e
        <p>« Le bruit qui se fait autour de cette affaire ne laisserait pas d'inquiéter le gouvernement provincial, qui, craignant tout
        <p>« On ne doute point en haut lieu que <persName>ce Messie</persName> dont le nom supposé est <persName>Aldavid</persName> ne s
        <ab type="ornament">⁎</ab>
        <p>Cette information, qui en son temps excita vivement la curiosité du public, m'incita, je ne sais pourquoi, à regretter l'abse
        <p>« Voilà une affaire propre à exciter l'imagination du baron, me disais-je. Il aurait sans doute bien des histoires <persName>
        <p>Et oubliant la <placeName>synagogue de Dollendorf</placeName>, je pensai à cet ami disparu, dont l'imagination et les habitud
        <ab type="ornament">⁎</ab>
        <p>Le lendemain, les journaux contenaient relativement à l'affaire de <placeName>Dollendorf</placeName> des informations plus se
        <p>Des dépêches, datées de <placeName>Francfort</placeName>, de <placeName>Mayence</placeName>, de <placeName>Leipzig</placeName
        <p>Comme à <placeName>Dollendorf</placeName>, il avait apparu devant une synagogue, la principale de chaque ville.</p>
        <p>La nouvelle s'étant vite répandue, <persName>les Juifs</persName> avaient accouru, et <persName>le Messie</persName> avait pr
        <p>À <placeName>Berlin</placeName>, vers cinq heures, la police ayant voulu s'emparer de lui, la foule juive, qui l'entourait, s
        <p>Pendant ce temps, <persName>Aldavid</persName> avait disparu comme par miracle…</p>
        <p>Ces nouvelles m'impressionnèrent, mais pas plus que le public qui se passionna pour <persName>Aldavid</persName>. Et, dans la
        <p>Partout l'émotion était à son comble et les gouvernements, comme on s'en souvient, tinrent des conseils dont les décisions fu
```

Position: Ln 1599, Ch 7 Total: Ln 1599, Ch 332182

☑ Toggle editor

Or load file from your computer

Aucun fichier sélectionné.

Select named-entity type

Places

REDEN parameters (optional, for experts)

Enter xpath expression (if apply)

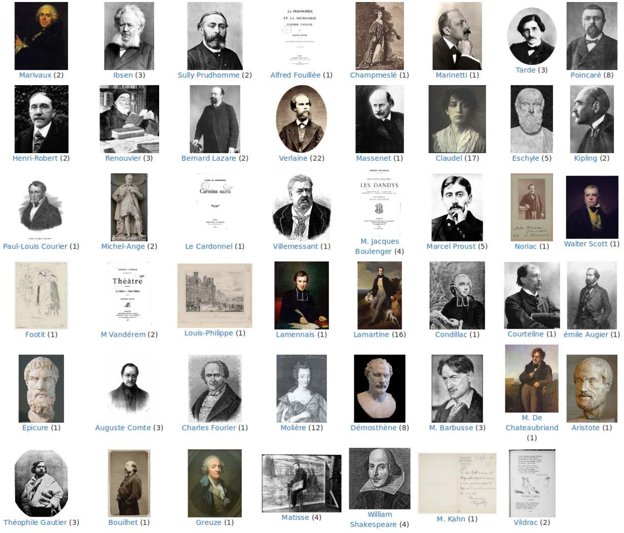


Figure 4 A visualization of authors mentioned in Thibaudet's "Réflexions sur la littérature", frequencies are displayed in parenthesis.

Bibliography

Brando, C., Frontini, F. and Ganascia, J. G. (2015). Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets. In Morzy, T., Valduriez, P. and Bellatreche, L. (Eds.), *New Trends in Databases and Information Systems*. (Communications in Computer and Information Science 539). Springer International Publishing, pp. 505–14.

Burnard, L. (2014). *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. (Encyclopédie Numérique). Marseille: OpenEdition Press.

Elliott, T. and Gillies, S. (2009). Digital geography and classics. *Digital Humanities Quarterly*, **3**(1).

Frontini, F., Brando, C. and Ganascia, J. G. (2015a). Domain-adapted named-entity linker using Linked Data. *Proceedings of the Workshop on NLP Applications: Completing the Puzzle*, vol. **1386**, Aachen: M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen. http://ceur-ws.org/Vol-1386/named_entity.pdf. (accessed 27 October 2015).

Frontini, F., Brando, C. and Ganascia, J. G. (2015b). Semantic Web based Named Entity Linking for Digital Humanities and Heritage Texts. *SW4SH 2015 Semantic Web for Scientific Heritage 2015*, vol. **1364**, Aachen: M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen, pp. 77–88, http://ceur-ws.org/Vol-1364/paper9.pdf.

Grassi, M., et al. (2012). Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries. *Proceedings of the 2nd International Workshop on Semantic Digital Archives, Paphos, Cyprus*, pp. 49–60.

Grassi, M., et al. (2013). Pundit: Augmenting Web Contents with Semantics. *Literary and Linguisting Computing*, **28**(4): 640–59.

Grossner, K., Janowicz, K. and Keßler, C. (2016). Place, Period, and Setting for Linked Data Gazetteers. In Mostern, Ruth, Berman, Lex and Southall, H. (Eds.), *Placing Names: Enriching and Integrating Gazetteers*. Bloomington, Indiana University Press http://geog.ucsb.edu/~jano/GrossnerJanowiczKessler_submitted_draft.pdf (accessed 27 October 2015).

Janowicz, K. (2009). The Role of Place for the Spatial Referencing of Heritage Data. *The Cultural Heritage of Historic European Cities and Public Participatory GIS Workshop*. The University of York, UK.

Jones, C. B., et al. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, **22**(10): 1045–65. doi:10.1080/13658810701850547.

Montuschi, P. and Benso, A. (2014). Augmented Reading: The Present and Future of Electronic Scientific Publications. *Computer*, **47**(1): 64–74 doi:10.1109/MC.2013.256.

Morbidoni, C., et al. (2013). Semantic Augmentation and Externalization in the Humanities: a Demonstrative Use Case. *Proceedings of the Digital Humanities 2013*, Lincoln, Nebraska.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso Books.

Mosallam, Y., Abi-Haidar, A. and Ganascia, J. G. (2014). Unsupervised Named Entity Recognition and Disambiguation: An Application to Old French Journals. *Advances in Data Mining. Applications and Theoretical Aspects*. Springer, pp. 12–23. http://link.springer.com/chapter/10.1007/978-3-319-08976-8_2 (accessed 27 July 2015).

Murrieta-Flores, P. and Gregory, I. (2015). Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology. *Open Archaeology*, **1**(1). doi:10.1515/opar-2015-0010. http://www.degruyter.com/view/j/opar.2014.1.issue-1/opar-2015-0010/opar-2015-0010.xml (accessed 27 October 2015).

Ramsay, S. (2008). Algorithmic Criticism. *Companion to Digital Literary Studies*. (Blackwell Companions to Literature and Culture). Oxford: Blackwell Publishing Professional http://www.digitalhumanities.org/companionDLS/ (accessed 24 February 2010).

Riguet, M. (in press). L'impact de la physiologie dans la critique littéraire de la fin du XIXe siècle : l'exemple de Claude Bernard, actes du colloque Littérature et Science au xixe siècle, dirigée par Elsa Courant et Romain Enriquez, ENS Ulm, avril 2015, Épistémocritique.

Schreibman, S. Digital Scholarly Editing. In Price, K. M. and Siemens, R. (eds), *Literary Studies in the Digital Age*. Modern Language Association of America http://dlsanthology.commons.mla.org/digital-scholarly-editing/ (accessed 5 March 2014).

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**(1): 3–26 doi:10.1075/li.30.1.03nad.

Stadler, C., et al. (2012). LinkedGeoData: A Core for a Web of Spatial Open Data. *Semantic Web Journal*, **3**(4): 333–54.

Hooland, S., et al. (2013). Exploring entity recognition and disambiguation for cultural heritage collections. *Digital*

Scholarship in the Humanities doi:10.1093/llc/fqt067. http://dsh.oxfordjournals.org/content/early/2014/12/02/llc.fqt067.

Notes

[1] REDEN is open source; you can find the code at https://github.com/cvbrandoe/REDEN.

[2] Find more information on OBVIL and ist digital library at http://obvil.paris-sorbonne.fr/.

New DH Publishing Models and Geopolitical Diversity

Isabel Galina Russell
igalina@unam.mx
Instituto de Investigaciones Bibliográficas -Universidad Nacional Autónoma de México - UNAM, Mexico

Ernesto Priani Saisó
epriani@gmail.com
Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México - UNAM, Mexico

The aim of this paper is to discuss how the new scholarly publishing models proposed by the Digital Humanities, if properly executed, may also serve to increase geopolitical diversity in the field. If not, they may well replicate vicissitudes of the traditional scholarly system and perpetuate the current deficiencies. The scholarly communication and publishing system currently in place is a highly complex and international structure that has over the past few decades come under increasing criticism as scholars debate its effectiveness, in particular in relation to new possibilities enabled by digital technologies. At the centre of these debates is the fact that scholarly publishing is both a communicative and collaborative practice that is vital for knowledge construction, as well as being an integral part of the academic reward system and therefore a essential part of both power and prestige within academia. It is directly linked to activities such as hiring, tenure, assignment of grant money, to name a few.

Researchers from periphery countries are sorely underrepresented in the global scholarly publishing system. They have relatively little participation in "international journals" where we find a dominance of publications from researchers in developed nations. Scientific production is measured in a number of 'core journals' that are determined by indexing services that tend to favour publications from certain regions of the world and published in English. This leads to the invisibility of research produced in periphery countries as "structural obstacles and subtle prejudices that prevent researchers in poor nations from sharing their discoveries with the industrial world and with each other" (Gibbs, 1995). The work being done in periphery countries does not participate on the global stage. Fiormonte (2015) analyzed Digital Humanities literature and found a strong predominance of citations to publications in English and articles about English speaking institutions and projects. In terms of knowledge construction it is important to note: "how the values of the Western intellect traditions are reflected in the conventions and practices of academic communities and their communications; how mainstream journals and their publishing practices are congenial to the interests of center knowledge while proving recalcitrant to periphery discourses; and how academic writing/publishing functions are an important means of legitimating and reproducing center knowledge" (Canagarajah 2002). As such we find that there is a marginalization of peripheries in the production of knowledge and the impact of the research.

The predominance of knowledge production from a handful of countries has important consequences, in particular the Humanities that require multilingual, multicultural heterogeneous environments if they are to fully represent the wide spectrum of human diversity. The Digital Humanities is a community that not only represents itself as collaborative and open (Spiro, 2012) but also sees itself as potentially transformative of the Humanities: "The tension between the digital humanities and the academic establishment is multifaceted and involves institutional hurdles to doing interdisciplinary and collaborative work, need for space and technological infrastructure, tenure systems not adapted to digital production and publications, and the need for non-faculty experts and corresponding career paths (Svensson, 2012). In this sense Digital Humanities in on the periphery of academia, seeking validation of the types of digital scholarship it is developing against established power structures and recognized and suitable forms of what is deemed 'valid scholarship'.

DH scholarship is produced in a variety of formats that are not necessarily monographs or journal articles that make up the traditional scholarly publishing system. Datasets, web pages, digital scholarly edition, textual markup and visualizations, to name a few, are part of DH production and the community has focused on how to certify them as valid outputs and forms of communicating knowledge. New forms of publishing, peer review and career paths are part of DH literature as it challenges the traditional ways of producing, disseminating, validating and certifying knowledge through the types of scholarly output it is producing.

For the Digital Humanities the possibility of new modes of communication and publishing have been fundamental in its construction. Initial work concentrated on digitizing and publishing texts online, what Davidson (2012) refers to as Humanities 1.0. The first attraction of online publishing is making available material that is of difficult access and/

or dispersed geographically (Priani, 2015). For others an important feature, with the relative low cost compared to publication on paper, was the possibility of making lesser-known materials available, such as non-canonical texts (Earhart 2012). From the periphery the possibility of electronic publishing offered a way of getting information published and noticed. Many Open Access projects, which focus on journal publishing, have worked towards this (Alperin, Fischman and Willinsky, 2008).

Since then however, it has become clear that the Internet provides the opportunity to change the way we think about publishing and what types of outputs can be considered valid forms of communicating knowledge. This in turn has led to discussion on how we can validate and certify this production. In order to change the system however, it requires "substantative rethinking (…) of the ways those faculty do their work, how they communicate that work, and how that work is read both inside and outside the academy" (Fitzpatrick, 2011). If we consider that the traditional scholarly publishing system has systematically excluded research from periphery countries, there is an opportunity, as we work towards new types of publishing and communication systems, to find ways of being deliberately inclusive. Although not referring specifically to periphery research Davidson (2012) idea of Humanities 2.0 which is "distinguished from monumental, first generation data-based projects not just by interactivity but also by an openness about participation grounded in a different set of theoretical premises, which decenter authority and knowledge", can be applied.

Many (Fiormonte, 2012; Liu, 2012; Rodríguez, 2012; Clavert, 2013; Dacos, 2013; Risam, 2015) have argued that DH must reflect more on the nature of the digital medium and the technologies that are being employed as well as addressing issues related to geo-linguistic diversity in the community. At the same time Digital Humanities is also a community about building and creating (Ramsay, 2011). If DH is indeed a transformative motor of academia, then reflecting from a critical perspective on the new types of digital scholarship that we are proposing is indispensable. We could propose new models that adequately incorporate digital scholarly output from countries on the periphery that are left out of the global publishing system within the traditional scholarly publishing model. If DH is proposing and fighting for new types of scholarly publishing, then should we not seek to build a model that takes this into consideration?

It is not possible of course to resolve this in a single conference presentation. The aim of this paper is to bring this subject to the table and to initiate a discussion in the different ways that this can be addressed. It is important to invest more in understanding the effects of the new types of publishing that we are advocating for as well as the digital infrastructures, primarily publishing platforms that we are developing and/or using. Discussing the implications of what we are building, the methods and structures we are using for communicating and publishing as well as the languages and materials we are prioritizing as part of the necessary self-reflection on what we are and what we do. As we advocate for new types of scholarship and we discuss new forms of peer review, certification, validation, publication and dissemination of these new types of publications we must make sure we do not incorporate tacit assumptions about the role and validity of periphery scholarship if not we shall inevitably continue to replication long-held prejudices and marginalization.

Bibliography

Alperin, J. P., Fischman, G. E. and Willinsky, J. (2008). Open access and scholarly publishing in Latin America: ten flavours and a few reflections. *Liinc em Revista*, **4**(2): 172-85.

Canagarajah, S. A. (2002). *A Geopolitics of Academic Writing.* USA: University of Pittsburgh Press.

Clavert, F. (2013). The DH Multicultural Revolution Did Not Happen Yet. *L'histoire contemporaine à l'ère numérique.*

Dacos, M. (2013). La estrategia de la sauna finlandesa. *Blog de la RedHD.*

Davidson, C. (2012). Humanities 2.0 - Promise, Perils, Predictions. In Gold, M. (ed.), *Debates in the Digital Humanities.* University of Minnesota Press.

Earhart, A. E. (2012). Can Information Be Unfettered? In Gold, M. K. (ed.). *Debates in the Digital Humanities.* University of Minnesota Press, pp. 309-18.

Fiormonte, D. (2012). Towards a Cultural Critique of the Digital Humanities. *Historical Social Research*, **37**: 59-76.

Gibbs, W. (1995). Lost Science in the Third World. *Scientific American,* pp. 92-99.

Fitzpatrick, K. (2011). *Planned Obsolescence: Publishing, Technology, and the Future of the Academy.* New York: New York University Press.

Fiormonte, D. (2015). Towards monocultural (digital) Humanities?, *InforLet.*

Liu, A. (2012). Where Is Cultural Criticism in the Digital Humanities? In Gold, M. K. (ed.). *Debates in the Digital Humanities*. University of Minnesota Press, pp. 495-98.

Priani, E. (2015). La Biblioteca Digital del Pensamiento Novohispano. *DHCommons*, 1.

Ramsay, S. (2011). On Building, *Stephan Ramsay Blog.*

Risam, R. (2015). Across Two (Imperial) Cultures, *Roopika Risam.*

Rodríguez Ortega, N. (2012). Prólogo: Humanidades Digitales y pensamiento crítico. In Romero Frías, E. e Sánchez González, M. (eds), *Ciencias Sociales y Humanidades Digitales.* CAC (Cuadernos Artesanos de Comunicación), pp.13-17.

Spiro, L. (2012). This is Why We Fight: Defining the Values of the Digital Humanities. In Gold, M. K. (ed.), *Debates in the Digital Humanities.* University of Minnesota Press, pp. 16-35.

Svensson, P.(2012). Envisioning the Digital Humanities. *DHQ*, **6**(1).

Crossed Semantic Analysis of Literary Texts with DeSeRT

Jean-Gabriel Ganascia
jean-gabriel.ganascia@lip6.fr
Sorbonne Universités, Université Pierre and Marie Curie (Paris 6); Laboratoire d'Informatique de Paris 6 (LIP6), CNRS UMR 7606; Labex OBVIL (Observatoire de la vie littéraire)

Chiara Mainardi
chiara85.mc@gmail.com
Sorbonne Universités, Université Paris-Sorbonne (Paris 4); Labex OBVIL (Observatoire de la vie littéraire)

Introduction

Doing comparative researches on a large literary corpus is often lengthy and demanding. With the digitization of contents, scholars have started using computers. Semantic information is essential for an appropriate understanding of texts and is an extremely important factor in cross-textual analysis. This is why we have developed a new semantic engine built especially for Digital Humanities, called DeSeRT[1]. This paper presents some of the results that were obtained from a corpus of the 17th and 18th century texts characteristic of the debates about theater in the classical age. The following paper is divided into four parts: after a first section that briefly describes the search engine used and the different investigations it allows, a second section introduces the corpus used. Then, the third section shows some of the results obtained.

The DeSeRT Search Engine

DeSeRT has been designed to identify and compare rewriting, paraphrasing or reformulation. It is based on an idea, already developed (Barron-Cedeño et al., 2013; Ferrero and Simac-Lejeune, 2015), according to which, even if the reformulations cannot be reduced to paraphrases, they retain the meaning of original texts by using either the same words or words of similar meaning. As a consequence, the detection of co-occurrences of a few semantically equivalent lemmas in small blocks of texts is sufficient to capture the equivalent meaning and, therefore, to identify the reformulation of the same ideas.

This is implemented in four steps:

1. Dividing texts into small blocks of words that are partially overlapping. Typically each block may contain 300 words and two consecutive blocks overlap by one-half, but both the block size and the proportion of overlap can vary.

2. Extracting the meaningful lemmas using a POS tagger. This step enables the exclusion of some syntactical categories, such as prepositions or articles, and to get the lemma associated to each word. The current implementation makes use of TreeTagger[2], but this could easily be changed.

3. Indexing each block with the lemmas obtained at step 3. Without going into implementation details, let us remark that the index is stored using the SQLite[3] database management system.

4. Retrieving blocks that contain many identical or semantically equivalent lemmas. This fourth module exploits a dictionary of synonyms to recover blocks that have similar meanings. It is also possible to use a thesaurus to restrict the search to a set of predefined terms, but this is not a necessity. The proximity between two blocks of text is based on an Okapi similarity measurement (Spärk-Jones et al, 2000), of which evaluation is greatly simplified by the fact that all blocks have the same size.

More precisely, the Okapi similarity measure of a block B with a set of terms T = *t*1, … *t*n is given by the following formula:

$$score(B,T) = \sum_{i=1}^{n} IDF(t_i).\frac{f(t_i,B).(k+1)}{f(t_i,B)+k.(1-b+b.\frac{|B|}{avdl})}$$

where *f*(t_i, *B*) is the frequency of the term t_i in the block *B*, |*B*| is the size of block *B*, *k* and *b* are two free parameters chosen in advance (usually chosen as k∈[1.2, 2] and b=0.75), *avdl* is the average document description length in the collection and *IDF*(t_i) is the *Inverse Document*

$$IDF(t_i) = Log(\frac{N-n(t_i)+0,5}{n(t_i)+0,5})$$

Frequency of term *t*i is given by:

It appears that, when the size of each block is the same and when the frequency of terms in blocks is supposed to be at max 1 (this approximation is justified since the blocks are small), this formula can be simplified. It then becomes equivalent to the information theoretic measure of the terms in blocks, i.e.

$$score(B,T) = \sum_{i=1}^{n} IDF(t_i) = \sum_{i=1}^{n} -\log(\Pr(t_i))$$

where *Pr*(t_i) is the probability of the term t_i in the overall corpus.

Using this score, it is possible to measure the similarity between blocks or between a block and a set *T* of terms t_i. As a consequence, DeSeRT can be used in different ways. The research queries may be done through words or concepts, i.e. words that are expanded using the dictionary of synonyms. It is also possible to compare any text (or file) to a corpus: then, DeSeRT detects corpus blocks where the meaning is similar to blocks of the given text, which allows, for instance, the arguments in a dispute or the anecdotes and the common places that are reused to be followed. Lastly, it is also possible (but not mandatory) to add a thesaurus or ontology to focus the search on a given semantic field.

Note that, based on techniques developed to detect plagiarism, many tools already exist that are designed to identify paraphrases, reuses and borrowings, i.e. sequences of words that are approximately identical, e.g. (Ganascia et al. 2014; Horson et al. 2010). However, these techniques are unable to spot reformulations of the same ideas or allusions to previous texts. DeSeRT has been designed to overcome these limitations.

The Hate of Theater

The project *Haine du Théâtre*[4] ("The Hatred of Theater" in French) aims to analyze theater disputes in Europe using scientific approaches and critical editions of polemical texts. The team's reflections are mainly focused around the discovery of the circumstances and the arguments used in theater controversies all across Europe, not limited to France, but also in England, Spain, Italy, and the Germanic area, from the last decades of the 16th century up to the beginning of the 19th century.

The corpus of the project collects many texts written in French during the 17th and the 18th centuries. The purpose of this project is to explore the gray areas of theater controversies in order to outline a global overview and to discover where and how the polemics began, their chronological discrepancies and the links between them and their contemporary resurgences. The total collection of the *Haine du Théâtre* texts is, by now, made up of 27 texts.

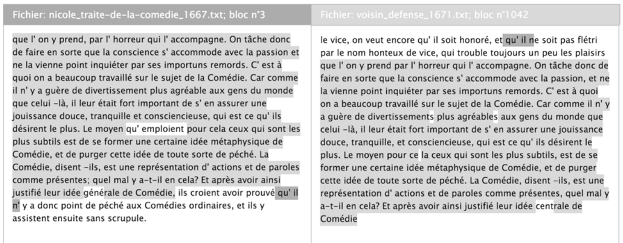
Exploitation of DeSeRT on the Hate of Theatre Corpus

Discovery of reuses

Querying the 27 texts of the *Haine du Théâtre* corpus, we found much reuse of similar passages and texts. DeSeRT is not only useful for detecting those parts of text that deal with the same concept, but is also a very good tool to find borrowings.

For example, comparing two texts, the *Défense* of Voisin (1666) and *Traité de la Comédie* of Nicole (1667), we discover immediately that the *Traité* has been included in the *Défense* by Voisin, which is a very long text, not only once, but twice. The first time, Voisin presents it as a re-publication, then he re-uses phrases similar to those employed by Nicole in different passages and he sprinkles them in his *Défense*.

The keywords of this correspondence are very well detected by DeSeRT, as can be seen in the example below.



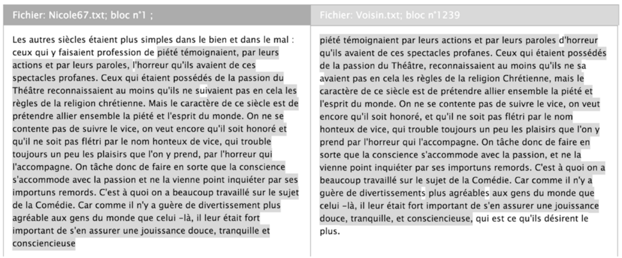
Fichier: nicole_traite-de-la-comedie_1667.txt; bloc n°3

que l' on y prend, par l' horreur qui l' accompagne. On tâche donc de faire en sorte que la conscience s' accommode avec la passion et ne la vienne point inquiéter par ses importuns remords. C' est à quoi on a beaucoup travaillé sur le sujet de la Comédie. Car comme il n' y a guère de divertissement plus agréable aux gens du monde que celui –là, il leur était fort important de s' en assurer une jouissance douce, tranquille et consciencieuse, qui est ce qu' ils désirent le plus. Le moyen qu' emploient pour cela ceux qui sont les plus subtils est de se former une certaine idée métaphysique de Comédie, et de purger cette idée de toute sorte de péché. La Comédie, disent –ils, est une représentation d' actions et de paroles comme présentes; quel mal y a–t–il en cela? Et après avoir ainsi justifié leur idée générale de Comédie, ils croient avoir prouvé qu' il n' y a donc point de péché aux Comédies ordinaires, et ils y assistent ensuite sans scrupule.

Fichier: voisin_defense_1671.txt; bloc n°1042

le vice, on veut encore qu' il soit honoré, et qu' il ne soit pas flétri par le nom honteux de vice, qui trouble toujours un peu les plaisirs que l' on y prend par l' horreur qui l' accompagne. On tâche donc de faire en sorte que la conscience s' accommode avec la passion, et ne la vienne point inquiéter par ses importuns remords. C' est à quoi on a beaucoup travaillé sur le sujet de la Comédie. Car comme il n' y a guère de divertissements plus agréables aux gens du monde que celui –là, il leur était fort important de s' en assurer une jouissance douce, tranquille, et consciencieuse, qui est ce qu' ils désirent le plus. Le moyen pour cela ceux qui sont les plus subtils, est de se former une certaine idée métaphysique de Comédie, et de purger cette idée de toute sorte de péché. La Comédie, disent –ils, est une représentation d' actions et de paroles comme présentes, quel mal y a–t–il en cela? Et après avoir ainsi justifié leur idée centrale de Comédie

Furthermore, continuing the analysis of the text we discover that in these two texts the actor is frequently associated with the idea of purity in religion.



Fichier: Nicole67.txt; bloc n°1 ;

Les autres siècles étaient plus simples dans le bien et dans le mal : ceux qui y faisaient profession de piété témoignaient, par leurs actions et par leurs paroles, l'horreur qu'ils avaient de ces spectacles profanes. Ceux qui étaient possédés de la passion du Théâtre reconnaissaient au moins qu'ils ne suivaient pas en cela les règles de la religion chrétienne. Mais le caractère de ce siècle est de prétendre allier ensemble la piété et l'esprit du monde. On ne se contente pas de suivre le vice, on veut encore qu'il soit honoré et qu'il ne soit pas flétri par le nom honteux de vice, qui trouble toujours un peu les plaisirs que l'on y prend, par l'horreur qui l'accompagne. On tâche donc de faire en sorte que la conscience s'accommode avec la passion et ne la vienne point inquiéter par ses importuns remords. C'est à quoi on a beaucoup travaillé sur le sujet de la Comédie. Car comme il n'y a guère de divertissement plus agréable aux gens du monde que celui –là, il leur était fort important de s'en assurer une jouissance douce, tranquille et consciencieuse

Fichier: Voisin.txt; bloc n°1239

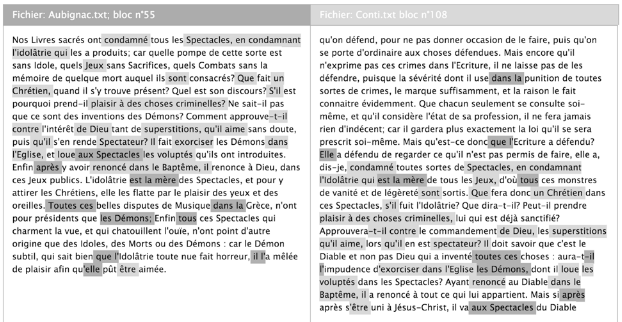
piété témoignaient par leurs actions et par leurs paroles d'horreur qu'ils avaient de ces spectacles profanes. Ceux qui étaient possédés de la passion du Théâtre, reconnaissaient au moins qu'ils ne sa avaient pas en cela les règles de la religion Chrétienne, mais le caractère de ce siècle est de prétendre allier ensemble la piété et l'esprit du monde. On ne se contente pas de suivre le vice, on veut encore qu'il soit honoré, et qu'il ne soit pas flétri par le nom honteux de vice, qui trouble toujours un peu les plaisirs que l'on y prend par l'horreur qui l'accompagne. On tâche donc de faire en sorte que la conscience s'accommode avec la passion, et ne la vienne point inquiéter par ses importuns remords. C'est à quoi on a beaucoup travaillé sur le sujet de la Comédie. Car comme il n'y a guère de divertissements plus agréables aux gens du monde que celui –là, il leur était fort important de s'en assurer une jouissance douce, tranquille, et consciencieuse, qui est ce qu'ils désirent le plus.

Reformulations of Ideas

DeSeRT also shows in detail the parts of the corpus that are similar or that develop the same ideas. This may either be done on demand, according user requests, i.e. to given texts, or to the overall corpus, which automates the process.

For instance, we have found many topics common not only to the texts by Nicole and Voisin, but also to two others e.g. the theme of the idolatry as the "mother of all spectacles" in (Aubignac, 1666) and (Conti, 1666).

Note that, in the following figures, we have greyed the identical passages with the MEDITE system, which only spots strict homologies, without considering many words that appear identical to DeSeRT because they correspond to the same lemma or to two synonymous lemmas. As a consequence, the number of gray zones considerably underestimates the semantic proximity detected by DeSeRT.

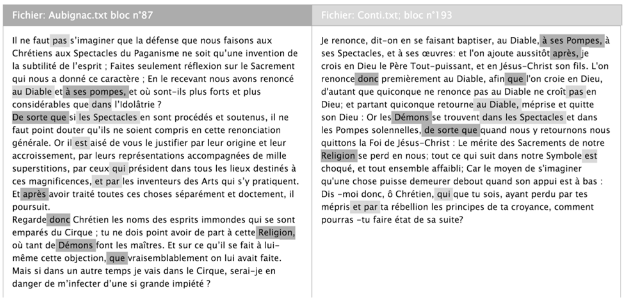


Fichier: Aubignac.txt; bloc n°55

Nos Livres sacrés ont condamné tous les Spectacles, en condamnant l'idolâtrie qui les a produits; car quelle pompe de cette sorte est sans Idole, quels Jeux sans Sacrifices, quels Combats sans la mémoire de quelque mort auquel ils sont consacrés? Que fait un Chrétien, quand il s'y trouve présent? Quel est son discours? S'il est pourquoi prend–il plaisir à des choses criminelles? Ne sait–il pas que ce sont des inventions des Démons? Comment approuve–t–il contre l'intérêt de Dieu tant de superstitions, qu'il aime sans doute, puis qu'il s'en rende Spectateur? Il fait exorciser les Démons dans l'Eglise, et loue aux Spectacles les voluptés qu'ils ont introduites. Enfin après y avoir renoncé dans le Baptême, il renonce à Dieu, dans ces Jeux publics. L'idolâtrie est la mère des Spectacles, et pour y attirer les Chrétiens, elle les flatte par le plaisir des yeux et des oreilles. Toutes ces belles disputes de Musique dans la Grèce, n'ont pour présidents que les Démons; Enfin tous ces Spectacles qui charment la vue, et qui chatouillent l'ouïe, n'ont point d'autre origine que des Idoles, des Morts ou des Démons : car le Démon subtil, qui sait bien que l'Idolâtrie toute nue fait horreur, il l'a mêlée de plaisir afin qu'elle pût être aimée.

Fichier: Conti.txt bloc n°108

qu'on défend, pour ne pas donner occasion de le faire, puis qu'on se porte d'ordinaire aux choses défendues. Mais encore qu'il n'exprime pas ces crimes dans l'Ecriture, il ne laisse pas de les défendre, puisque la sévérité dont il use dans la punition de toutes sortes de crimes, le marque suffisamment, et la raison le fait connaitre évidemment. Que chacun seulement se consulte soi–même, et qu'il considère l'état de sa profession, il ne fera jamais rien d'indécent; car il gardera plus exactement la loi qu'il se sera prescrit soi–même. Mais qu'est–ce donc que l'Ecriture a défendu? Elle a défendu de regarder ce qu'il n'est pas permis de faire, elle a, dis–je, condamné toutes sortes de Spectacles, en condamnant l'Idolâtrie qui est la mère de tous les Jeux, d'où tous ces monstres de vanité et de légèreté sont sortis. Que fera donc un Chrétien dans ces Spectacles, s'il fuit l'Idolâtrie? Que dira–t–il? Peut–il prendre plaisir à des choses criminelles, lui qui est déjà sanctifié? Approuvera–t–il contre le commandement de Dieu, les superstitions qu'il aime, lors qu'il en est spectateur? Il doit savoir que c'est le Diable et non pas Dieu qui a inventé toutes ces choses : aura–t–il l'impudence d'exorciser dans l'Eglise les Démons, dont il loue les voluptés dans les Spectacles? Ayant renoncé au Diable dans le Baptême, il a renoncé à tout ce qui lui appartient. Mais si après après s'être uni à Jésus–Christ, il va aux Spectacles du Diable



Fichier: Aubignac.txt bloc n°87

Il ne faut pas s'imaginer que la défense que nous faisons aux Chrétiens aux Spectacles du Paganisme ne soit qu'une invention de la subtilité de l'esprit ; Faites seulement réflexion sur le Sacrement qui nous a donné ce caractère ; En le recevant nous avons renoncé au Diable et à ses pompes, et où sont–ils plus forts et plus considérables que dans l'Idolâtrie ?

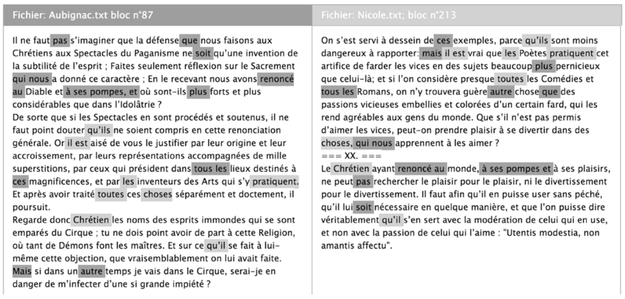
De sorte que si les Spectacles en sont procédés et soutenus, il ne faut point douter qu'ils ne soient compris en cette renonciation générale. Or il est aisé de vous le justifier par leur origine et leur accroissement, par leurs représentations accompagnées de mille superstitions, par ceux qui président dans tous les lieux destinés à ces magnificences, et par les inventeurs des Arts qui s'y pratiquent. Et après avoir traité toutes ces choses séparément et doctement, il poursuit.

Regarde donc Chrétien les noms des esprits immondes qui se sont emparés du Cirque ; tu ne dois point avoir de part à cette Religion, où tant de Démons font les maîtres. Et sur ce qu'il se fait à lui–même cette objection, que vraisemblablement on lui avait faite. Mais si dans un autre temps je vais dans le Cirque, serai–je en danger de m'infecter d'une si grande impiété ?

Fichier: Conti.txt; bloc n°193

Je renonce, dit–on en se faisant baptiser, au Diable, à ses Pompes, à ses Spectacles, et à ses œuvres: et l'on ajoute aussitôt après, je crois en Dieu le Père Tout–puissant, et en Jésus–Christ son fils. L'on renonce donc premièrement au Diable, afin que l'on croie en Dieu, d'autant que quiconque ne renonce pas au Diable ne croît pas en Dieu; et partant quiconque retourne au Diable, méprise et quitte son Dieu : Or les Démons se trouvent dans les Spectacles et dans les Pompes solennelles, de sorte que quand nous y retournons nous quittons la Foi de Jésus–Christ : Le mérite des Sacrements de notre Religion se perd en nous; tout ce qui suit dans notre Symbole est choqué, et tout ensemble affaibli; Car le moyen de s'imaginer qu'une chose puisse demeurer debout quand son appui est à bas : Dis –moi donc, ô Chrétien, qui que tu sois, ayant perdu par tes mépris et par ta rébellion les principes de ta croyance, comment pourras –tu faire état de sa suite?

Secondly, as shown in the following figures, the topic "renouncing the Devil" (*Renoncer au Diable* in French) is regularly present in the texts written by Aubignac, Conti and Voisin,[5] while it appears only once in (Nicole, 1667: 477).



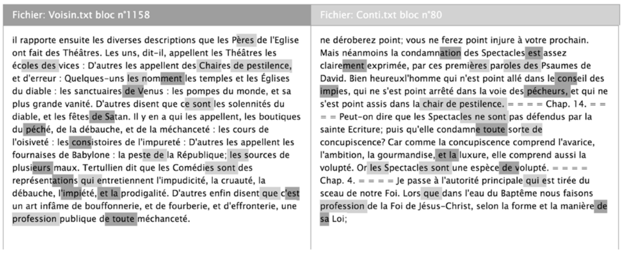
Fichier: Aubignac.txt bloc n°87

Il ne faut pas s'imaginer que la défense que nous faisons aux Chrétiens aux Spectacles du Paganisme ne soit qu'une invention de la subtilité de l'esprit ; Faites seulement réflexion sur le Sacrement qui nous a donné ce caractère ; En le recevant nous avons renoncé au Diable et à ses pompes, et où sont-ils plus forts et plus considérables que dans l'Idolâtrie ?
De sorte que si les Spectacles en sont procédés et soutenus, il ne faut point douter qu'ils ne soient compris en cette renonciation générale. Or il est aisé de vous le justifier par leur origine et leur accroissement, par leurs représentations accompagnées de mille superstitions, par ceux qui président dans tous les lieux destinés à ces magnificences, et par les inventeurs des Arts qui s'y pratiquent. Et après avoir traité toutes ces choses séparément et doctement, il poursuit.
Regarde donc Chrétien les noms des esprits immondes qui se sont emparés du Cirque ; tu ne dois point avoir de part à cette Religion, où tant de Démons font les maîtres. Et sur ce qu'il se fait à lui-même cette objection, que vraisemblablement on lui avait faite. Mais si dans un autre temps je vais dans le Cirque, serai-je en danger de m'infecter d'une si grande impiété ?

Fichier: Nicole.txt; bloc n°213

On s'est servi à dessein de ces exemples, parce qu'ils sont moins dangereux à rapporter: mais il est vrai que les Poètes pratiquent cet artifice de farder les vices en des sujets beaucoup plus pernicieux que celui-là; et si l'on considère presque toutes les Comédies et tous les Romans, on n'y trouvera guère autre chose que des passions vicieuses embellies et colorées d'un certain fard, qui les rend agréables aux gens du monde. Que s'il n'est pas permis d'aimer les vices, peut-on prendre plaisir à se divertir dans des choses, qui nous apprennent à les aimer ?
=== XX. ===
Le Chrétien ayant renoncé au monde, à ses pompes et à ses plaisirs, ne peut pas rechercher le plaisir pour le plaisir, ni le divertissement pour le divertissement. Il faut afin qu'il en puisse user sans péché, qu'il lui soit nécessaire en quelque manière, et que l'on puisse dire véritablement qu'il s'en sert avec la modération de celui qui en use, et non avec la passion de celui qui l'aime : "Utentis modestia, non amantis affectu".

Further results of DeSeRT lead us to understand that many others expressions are common to the four authors, such as the description of the theatre as a "flesh of pestilence" (*chair de pestilence*) or as a "school of the debauchery".

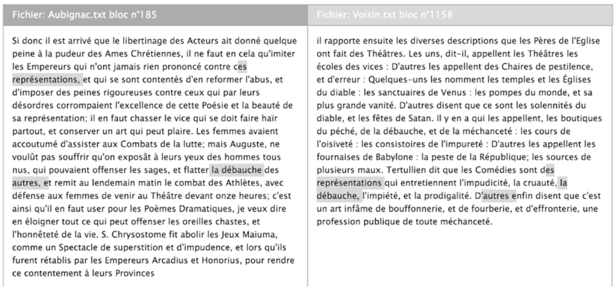


Fichier: Voisin.txt bloc n°1158

il rapporte ensuite les diverses descriptions que les Pères de l'Eglise ont fait des Théâtres. Les uns, dit-il, appellent les Théâtres les écoles des vices : D'autres les appellent des Chaires de pestilence, et d'erreur : Quelques-uns les nomment les temples et les Églises du diable : les sanctuaires de Venus : les pompes du monde, et sa plus grande vanité. D'autres disent que ce sont les solennités du diable, et les fêtes de Satan. Il y en a qui les appellent, les boutiques du péché, de la débauche, et de la méchanceté : les cours de l'oisiveté : les consistoires de l'impureté : D'autres les appellent les fournaises de Babylone : la peste de la République; les sources de plusieurs maux. Tertullien dit que les Comédies sont des représentations qui entretiennent l'impudicité, la cruauté, la débauche, l'impiété, et la prodigalité. D'autres enfin disent que c'est un art infâme de bouffonnerie, et de fourberie, et d'effronterie, une profession publique de toute méchanceté.

Fichier: Conti.txt bloc n°80

ne déroberez point; vous ne ferez point injure à votre prochain. Mais néanmoins la condamnation des Spectacles est assez clairement exprimée, par ces premières paroles des Psaumes de David. Bien heureuxl'homme qui n'est point allé dans le conseil des impies, qui ne s'est point arrêté dans la voie des pécheurs, et qui ne s'est point assis dans la chair de pestilence. = = = = Chap. 14. = = = = Peut-on dire que les Spectacles ne sont pas défendus par la sainte Ecriture; puis qu'elle condamne toute sorte de concupiscence? Car comme la concupiscence comprend l'avarice, l'ambition, la gourmandise, et la luxure, elle comprend aussi la volupté. Or les Spectacles sont une espèce de volupté. = = = = Chap. 4. = = = = Je passe à l'autorité principale qui est tirée du sceau de notre Foi. Lors que dans l'eau du Baptême nous faisons profession de la Foi de Jésus-Christ, selon la forme et la manière de sa Loi;



Fichier: Aubignac.txt bloc n°185

Si donc il est arrivé que le libertinage des Acteurs ait donné quelque peine à la pudeur des Ames Chrétiennes, il ne faut en cela qu'imiter les Empereurs qui n'ont jamais rien prononcé contre ces représentations, et qui se sont contentés d'en reformer l'abus, et d'imposer des peines rigoureuses contre ceux qui par leurs désordres corrompaient l'excellence de cette Poésie et la beauté de sa représentation; il en faut chasser le vice qui se doit faire haïr partout, et conserver un art qui peut plaire. Les femmes avaient accoutumé d'assister aux Combats de la lutte; mais Auguste, ne voulût pas souffrir qu'on exposât à leurs yeux des hommes tous nus, qui pouvaient offenser les sages, et flatter la débauche des autres, et remit au lendemain matin le combat des Athlètes, avec défense aux femmes de venir au Théâtre devant onze heures; c'est ainsi qu'il en faut user pour les Poèmes Dramatiques, je veux dire en éloigner tout ce qui peut offenser les oreilles chastes, et l'honnêteté de la vie. S. Chrysostome fit abolir les Jeux Maiuma, comme un Spectacle de superstition et d'impudence, et lors qu'ils furent rétablis par les Empereurs Arcadius et Honorius, pour rendre ce contentement à leurs Provinces

Fichier: Voisin.txt bloc n°1158

il rapporte ensuite les diverses descriptions que les Pères de l'Eglise ont fait des Théâtres. Les uns, dit-il, appellent les Théâtres les écoles des vices : D'autres les appellent des Chaires de pestilence, et d'erreur : Quelques-uns les nomment les temples et les Églises du diable : les sanctuaires de Venus : les pompes du monde, et sa plus grande vanité. D'autres disent que ce sont les solennités du diable, et les fêtes de Satan. Il y en a qui les appellent, les boutiques du péché, de la débauche, et de la méchanceté : les cours de l'oisiveté : les consistoires de l'impureté : D'autres les appellent les fournaises de Babylone : la peste de la République; les sources de plusieurs maux. Tertullien dit que les Comédies sont des représentations qui entretiennent l'impudicité, la cruauté, la débauche, l'impiété, et la prodigalité. D'autres enfin disent que c'est un art infâme de bouffonnerie, et de fourberie, et d'effronterie, une profession publique de toute méchanceté.

Conclusion

To conclude, DeSeRT allows the discovery of reformulations and similar phrases, as well as related topics and passages in a corpus. It is usable on any kind of corpus that is made up by files in the txt format. As we have briefly shown in this paper, this search engine enables users to identify crucial passages of a specific corpus according to two types of detection: the discovery of reuses, such as plagiarism or hidden rewritings, and the reformulation of ideas, which can be manually given by the user (as words or concepts) or automatically extracted by the DeSeRT search engine.

Bibliography

Barron-Cedeño, A., et al. (2013). Plagiarism meets paraphrasing : Insights for the next generation in automatic plagiarism detection. In *Association for Computational Linguistics*, vol. **39**: 917–47.

Conti, Prince de, A. de B. (1666). *Traité de la Comédie et des spectacles*, Louis Billaine, Paris.

D'Aubignac, Abbé, F.H. (1666). *Dissertation sur la condemnation des théâtres*, N. Pépingué, Paris.

Ferrero, J. and Simac-Lejeune, A. (2015). Détection automatique de reformulations - Correspondance de concepts appliquée à la détection du plagiat, *15ème conférence internationale sur l'extraction et la gestion des connaissances (EGC 2015)*, Luxembourg.

Ganascia, J.G., Glaudes, P. and Del Lungo, A. (2014). Automatic detection of reuses and citations in literary texts, *Literary and Linguistic Computing*, 2014, doi: 10.1093/llc/fqu020

Horton, R., Olsen, M., and Roe, G. (2010). Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections, *Digital Studies/ Le champ numérique* **2**(1), Available at: http://www.digitalstudies.org/ojs/index.php/ digital_studies/article/view/190/235. (last access 7 November 2013).

Nicole, P. (1667). *De la Comédie*, Adolphe Beyers, Liege.

Spärck Jones, K., Walker, S. and Robertson, S. E. (2000). *A probabilistic model of information retrieval: Development and comparative experiments: Part 1.* Information Processing and Management **36** (6): 779–808.

Voisin, Abbé, J. (1671). *Défense du traité de Mgr le Prince de Conti touchant la comédie et les spectacles ou la réfutation d'un livre intitulé Dissertation sur la condamnation des théâtres*, Coignard, Paris.

Notes

[1] A French version of DeSeRT is freely available online at http://obvil-dev.paris-sorbonne.fr/desert/

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[3] http://sqlite.org/

[4] http://obvil.paris-sorbonne.fr/projets/la-haine-du-theatre

[5] D'Aubignac, 1666 : 59, 62, 65, 72, 73, 74, 79, 217. Conti, 1666 : 88, 105, 120, 144, 173, 182, 184. Voisin, 1671 : 59, 86, 88, 97, 113, 114, 124, 165, 205, 212, 228, 407, 427, 433, 451, 463, 481.

Pulp Science Fiction's Legacy to Women in Science

Elizabeth Winfree Garbee
egarbee@asu.edu
Arizona State University, United States of America

This study examines a "pulp" science fiction corpus (1930–1965) through corpus linguistic analysis in order to digitally reconstruct the gendered occupational identities created by those authors, and the culture they represent, which perpetuated a stereotype of "the scientist" and how they characterized women in professional scientific roles. I created "occupational archetypes" based on the linguistic analysis of collocates, clusters, and textual examples of science, technology, engineering, and math (STEM) career keywords in order to investigate the culturally informed gender roles demonstrated in modern stereotypes of the scientist. I chose to study pulp science fiction, a sub-genre of science fiction literature that enjoyed a wide readership during the formative decades of the creation of the scientific industrial complex in pre and post-war America. One way to get at the culture that created and then maintained our national scientific industrial complex is through examining the stories about science that people of that time produced. For indeed, stories, and even more simply, language are a transmitter of social and cultural values, especially when it comes to gender roles (Rey, 2001). Pulp science fiction existed as a sub-genre of science fiction from roughly 1930 to 1965, characterized by its wide audience and affordability. The accessibility and engaging style of this literary genre gained it a wide readership, and "the pulps" as they came to be called quickly became a feature of American life during the pre and post-war eras. These stories, and indeed the genre at large, represent popular conceptions of "appropriate" gender identities and reinforce those occupational stereotypes that play such a key role in the lives of women scientists.

The aim of corpus linguistics is to study patterns of language at their most fundamental level of words and phrases, thereby revealing patterns of meaning, making the implicit explicit (Biber, 1998; Biber, 2009; Stubbs, 2001; McEnery, 2001; Lakoff, 2008; Kennedy, 2014; Hettel, 2013). Since patterns of meaning are precisely what I wished to investigate with respect to gendered occupational stereotypes, this method served as the basis for my study. Corpus linguistics is able to harness the power of Moretti's distant reading approach (Moretti, 2013) in uncovering the scope and and nature of the literature, while also providing clues as to which specific pieces within a corpus merit a close reading.

The corpus I constructed for this project consists of 560 full text copies of pulp science fiction stories from 1930 to 1965 (totaling just over 6 million words), published in magazines like Astounding Stories, Amazing Stories, Analog Science Fact and Fiction, Planet Stories, and If Worlds of Science Fiction. These full texts were obtained from public repositories, principally Project Gutenberg (https://www.gutenberg.org/) and The Internet Archive's Pulp Magazine Archive (https://archive.org/details/pulpmagazinearchive). While some of these stories were already conveniently in plain text files, others were scanned copies of the original pulp magazine pages stored as image files. The latter I converted into plain text through the application of Tesseract, an open source optical character recognition (OCR) program. I then organized these stories according to their date of publication in the magazines, stratifying according to five year periods: 1930-34, 1935-39, etc. Each of these five year periods contain 80 stories, coming to 560 in total. This stratification allows for representativeness through ensuring that all five year periods were weighted proportionally over time (Sinclair, 2004).

When I finished constructing the corpus, I used the software suite WordSmith Tools to generate keyword lists for the corpus in its entirety, in addition to each five year period respectively. When I generated the keyword lists for each five year period, I used the rest of the pulp science fiction corpus as my reference, in order to track how these words were being used over time (Bondi, 2010). From these general keyword lists, I chose the keywords which represented careers or occupations that constitute or interact with STEM disciplines: scientist, engineer, mathematician, doctor, nurse, and professor. I then used WordSmith Tools to analyze measures of association for the above science, technology, and engineering occupational words. By focusing on the language used to describe occupations related to the sciences, I was able to get a picture of the characterization of these professions at the time the stories were published. I also did a collocation analysis in order to uncover the words that most frequently co-occurred with these keywords, limited to five words to the right and left of the key word in question (the node). The character of the collocates reflects the nature of the node, and the distinctions offered by collocations are subtle, yet crucial to the creation of a linguistic profile (Hettel, 2013).

In order to determine which collocates were statistically significant, I evaluated the association by its t-score, a statistic which works well with smaller corpora (such as mine) because it also takes frequencies into account, as opposed to mutual information (MI). Using the STEM occupation keywords, collocates, clusters, and qualitative examination of specific examples of the node words in context, I then created "lexical profiles" of each of these science, technology, engineering, and related occupations, which I'm terming "occupational archetypes." The development of these archetypes is based largely on the work done by Hettel on the construction of lexical profiles from collocations, clusters, and context in the language of US nuclear plants and regulatory entities.

Specifically, the archetype I constructed of "the scientist" revealed a middle aged white male, defined by his adherence to "true" or "good" science, and often called upon by other characters to provide scientific or technical insight. Though he spends a good deal of his time talking, others struggle to grasp his meaning and find him difficult to deal with. This occupational archetype is a mirror image of the American stereotype of scientific professionals, one which leaves no room for diversity in race or gender. Furthermore, through this analysis, I discovered that out of the hundreds of scientists in my corpus, only three were women. A linguistic analysis of these women in particular (a chemist, a physicist, and a mathematician) revealed American cultural assumptions about the intersection of femininity and science: a female scientist could either be beautiful or accomplished. And even then, the chemist's beauty came with exploitation (and the physicist's ugliness with prestige), and scientific genius in these women necessitated qualification (i.e. genius "in her own way") while that of their male colleagues did not. The mathematician, the one woman in the corpus with beauty and brains, so to speak, appeared very late on the scene, and perhaps signals a shift in the cultural conception of who a scientist could be and what they could look like. Making these entrenched cultural stereotypes of women in science explicit through linguistic analysis is the first step in creating a STEM workforce strong through its diversity and acceptance.

Bibliography

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Bondi, M. and Scott, M. (eds.) (2010). Keyness in texts. *John Benjamins Publishing*, vol. **41**.

Gries, S. T. (2010). Useful statistics for corpus linguistics. *A mosaic of corpus linguistics: selected approaches*, pp. 269-91.

Hettel, J. M. (2013). *Harnessing the power of context*.

Kennedy, G. (2014). *An introduction to corpus linguistics*. Routledge.

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

McEnery, T. and Wilson, A. (2001). Corpus linguistics: An introduction. Edinburgh University Press.

Moretti, F. (2013). *Distant reading. Verso Books.*

Oakes, Michael P. (2010). *Statistics for Corpus Linguistics*. Edinburgh University Press. 1998.

Rey, J. M. (2001). Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966-1993. *Variation in English: Multidimensional Studies*. London: Longman. pp. 13856.

Sinclair, J. (2004). Developing linguistic corpora: a guide to good practice. *Corpus and text–basic principles.*

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell Publishers.

Tognini-Bonelli, E. (2001). Corpus linguistics at work,*John Benjamins Publishing*, vol. **6**.

Building Blocks of Fiction: Lexical Bundles in Nineteenth-Century Novels

Marissa Lynn Gemma
marissa.gemma@gmail.com
Max Planck Institute for Empirical Aesthetics

Ryan James Heuser
heuser@stanford.edu
Stanford University

Introduction

Several decades ago, corpus linguists undertook systematic analyses of a lexical unit larger than the individual word but smaller than syntactical or phrasal units—multiword sequences they called "lexical bundles" (Biber et al., 1999: 989; Biber and Conrad 1999: 58). Lexical bundles are extremely common collocations of three or more words, such as "*I want to*", "*it was a*", or "*going to be a.*" Unlike idioms or clichés, which are statistically very rare—and unlike word n-grams, which can occur with any frequency— lexical bundles are diffused throughout the language, occurring at least ten times per million words, and in many cases much more frequently (Biber et al., 1999: 989).[1] Because of their frequency, lexical bundles are regarded as "discourse framing devices": sequences of words that function as connective tissue in organizing discourse, expressing stance, or conveying referential status (Biber, 2006: 174).

While much of this ground-breaking research has focused on register—i.e., on differences in the use of multi-word sequences in various discourse contexts (like conversation vs. academic prose)—surprisingly little attention has been paid to the role of bundles in fiction. Biber et al. (1999) do not discuss results for fiction in their work, focusing instead on differences between academic prose and conversation (cf. also Conrad and Biber, 2004). Michaela Mahlberg's corpus stylistics approach to Dickens' fiction (2012) was pioneering in the use of lexical bundles to study fiction, but to date, no studies focusing on corpus- or register-wide trends in fiction have appeared.[2] Yet the dramatic differences in the use of lexical bundles in other registers suggest that lexical bundles play an important role in building the discourse of fiction, and also that the bundles that occur in fiction will be significantly different than those occurring in other registers.

This paper analyzes the role of lexical bundles in a corpus of over 1000 novels published in Great Britain and America between 1800 and 1905. The two key contributions of the paper are: 1) to provide a taxonomy of the discourse functions of lexical bundles in nineteenth-century English and American fiction; and 2) to historicize that usage by tracking changes in our corpus over the course of the nineteenth century. We also provide some data about differences between narration and dialogue.

In expanding the unit of analysis beyond the level of the word, this paper also aims to intervene in recent methodological debates about digital humanities research on literary style. Much DH research on style in recent years—including some of the authors' own work—has relied on a bag-of-words approach (Heuser and Le-Khac, 2012; Underwood and Sellers, 2012; for a discussion of the virtues of this approach, see Underwood, 2013). Since the terrain of higher-level lexical patterning in fiction remains under-explored, this paper contributes to the field both a methodological approach and a set of empirical results about the language of the nineteenth-century Anglo-American novel.

Methods

Our corpus derives from two digital fiction collections (licensed to the Stanford University Libraries) by Proquest: "Early American Fiction" (805 American novels published from 1789 - 1875; and "Nineteenth-Century Fiction" (250 British novels published from 1782-1903). To extend the American corpus' historical scope to match that of the British corpus, we added a collection of about 325 American novels published between 1875 and 1905. We selected these texts based on their inclusion in the *Annals of American Literature* (Ludwig and Nault, 1989) and their availability in Project Gutenberg.

To identify bundles in a corpus with an uneven historical distribution of texts, we split each national corpus into four twenty-five year segments, creating eight sub-corpora. Each sub-corpus is derived from an identical number of words per author. Authors with fewer than 100,000 words in a particular period were not included; authors with more than this number were included by selecting 100 random slices of 1,000 words from all of their texts published in the period. Sub-corpora ranged in length from 900,000 words (U.S. publications, 1800-25) to 5.9 million words (U.S. publications, 1850-75), with a median length of 2.3 million words.

Following Biber et al. (1999), we defined lexical bundles as the most commonly occurring tri- and quad-grams in our sub-corpora, with a threshold of at least ten occurrences per one million words (frequency per million [FPM]). After tokenizing each of the eight sub-corpora, we counted the number of occurrences of each unique tri-gram and quad-gram, normalizing by the length of the sub-corpus.[3] Any tri-gram or quad-gram with a frequency above 10 FPM in any of the eight sub-corpora was considered a potential lexical bundle. Biber et al. additionally required that bundles occur in at least five different texts in the corpus, in order to guard against the possibility of individual authorial or textual effects (991). Similarly, we excluded from our list of bundles those that occurred in fewer than three unique authors, so as to exclude idiosyncratic stylistic habits, as well as bundles containing character names or other novel-specific traits.[4]

We manually created an interpretive typology of the most frequent lexical bundles: the 150 most frequent tri-gram bundles (all with a median FPM across sub-corpora above 31), and the 240 most frequent quad-gram bundles (all with a median FPM across sub-corpora above 13). By looking at randomly-selected examples from the sub-corpora, each of these lexical bundles was annotated for its apparent function within fictional discourse. For example, "there was a" (5th most frequent tri-gram bundle, with a median FPM of 184) was annotated as "expletive": grammatically, "expletives" are phrases of the form ["there" or "it" + to be], and within fiction, they provide a means by which the existence or effect of something can be easily introduced.[5]

Finally, because corpus linguistics research has shown dramatic differences in lexical bundles across oral and written registers, we separately tracked their frequency in the narration and dialogue portions of our texts. To separate dialogue from narration, we used a tool developed by Grace Muzny at Stanford University. For this task we used a slightly reduced version of our corpus, which we curated by hand to ensure proper typographic markings of dialogue and thus a high precision and recall for the dialogue separation. We then replicated the corpus design described above, creating eight sub-corpora for each register, dialogue and narration. In this case, however, the periods are of twenty-year increments, from 1825 to 1905, due to the paucity of typographically well-formatted novels previous to 1825; also, each author contributes not 100,000 but 50,000 words to a particular sub-corpus, and authors with fewer than this number are not included in the sub-corpus.

Results

Due to constraints of space, we can provide here only an overview of our findings in each area.

1. The function of lexical bundles in nineteenth-century fiction

From our annotations (see "Methods" above), the most frequently-occurring bundles in fiction have one of the following functions:

- Expletive (*there was a, it was a, there was no, it was not, and there was*)
- Auxiliary forms (*i do not, i did n't, he did not, he had been*)
- Modal forms (*i could not, would have been, i ca n't, as if he*)
- Relative clause markers (*that he had, that he was, which he was, that she was*)
- Temporal markers (*for a moment, as soon as, the first time*)
- Partitive constructions (*one of the, part of the, some of the*)
- Spatial markers (*out of the, at the door, in the house*)
- Stance markers (*do n't know, i am sure, to be sure, seemed to be*)
- *Of*-genitive (*the name of; the voice of; the heart of; the hands of*)
- Discourse organization markers (*in spite of, as to the, in order to*)

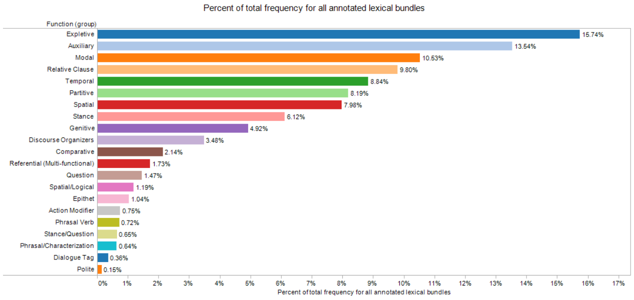


Figure 1: The most frequent 240 4-gram bundles, and most frequent 150 3-gram bundles, were annotated for their primary function in fiction; displayed here is the percentage of occurrences of all 390 annotated bundles within each unique functional type

The discursive mechanisms and requirements of fiction are immediately discernable in this ranked typology, especially when compared to prior work on lexical bundles in other discursive contexts. The prominence of expletives in fiction, for example, seems particularly significant, as they condense a fundamental gesture of storytelling into a phrase: once upon a time, *there was a*... Expletives function in our corpus as a means of positing and sustaining a fictional ontology: *it was* the best of times, *it was* the worst of times. Similarly, the prevalence of temporal and spatial markers (*as soon as, in front of*) help perform another fundamental task of fictional narration, that is, maintaining a complex and evolving network of persons, objects and events in their changing relationships to one another.

2. Historical behavior of lexical bundles

On the whole, the historical trends we find in the use of lexical bundles suggest an increasing specialization of certain narrative functions over the course of the century, especially in the language of dialogue and its narrative orchestration. For example, the most frequent functional type of lexical bundle—the "expletive" bundles described above, like " *there was a*"—are actually more frequent in dialogue than in narration earlier in the century, but are then increasingly adopted by fictional narrators as part of their machinery for coordinating the existence of fictional objects (see Figure 2).

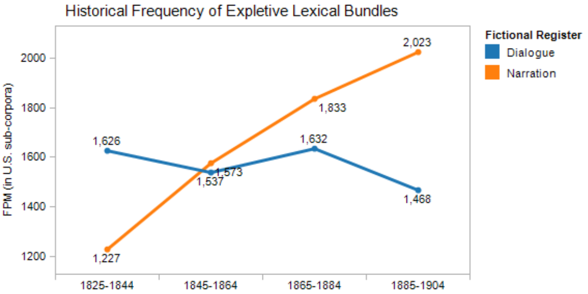


Figure 2: The sum frequency per million of 21 lexical bundles annotated as expletives. Frequency is calculated in the dialogue and narrative portions, per period, of U.S. fiction

We also see evidence of historical changes that do not depend on internal register differences in fiction (narration vs. dialogue). Particularly striking is the general decline in the use of *of*-genitives in both narration and dialogue, in both British and American fiction (see Figure 3)—presumably replaced by the more concise Saxon genitive (*'s*) or by noun phrases (as in *the center of the city* vs. *the city center*). This may be an instance of a broader shift towards more informal, concise diction as the nineteenth century progresses, since the Saxon genitive and such noun phrases privilege efficiency of expression over the rhythmical advantages of the *of*-genitive.

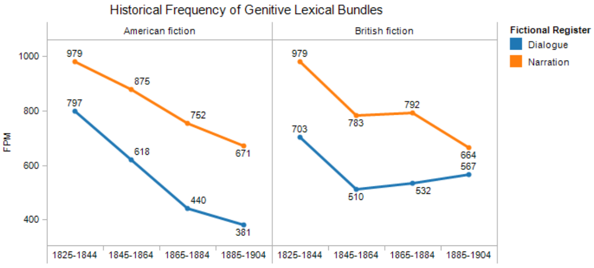


Figure 3: The sum frequency per million for 35 lexical bundles annotated as genitive, per period, in the dialogue and narrative portions of U.S. and British fiction

Finally, we find some key differences in the use of bundles between dialogue and narration, largely having to do with national differences in colloquial expressions. In American dialogue, for example, over the course of the century we find an increased use of *going-to* future forms (*are going to; am going to; going to do; are you going; going to be*), along with an increase in colloquial discourse markers (*sort of thing; kind of a; all the time*). In the British dialogue corpus, similarly, we find a decrease in formal and polite phrases over time (*to be sure; i have no doubt; my dear sir; depend upon it; god bless you; as you please;*

by the bye; the honour of), and a concurrent rise in more informal modern phrases (*in spite of; at any rate; here and there; now and again*). Taken together, such trends suggest that this shift towards more informal bundles is particularly concentrated in fictional dialogue.

Bibliography

Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In *Out of Corpora: Studies in Honor of Stig Johansson*, (Ed) Hilde Hasselard, Signe Oksefjell. Rodopi, Amsterdam: Rodopi, pp: 181–89.

Biber, D., et al. (2003). Lexical bundles in speech and writing: An initial taxonomy. In Wilson, A. et al., *Corpus Linguistics by the Lune*. Frankfurt am Main: Peter Lang, pp. 71–92.

Biber, D., et. al. (1999). *The Longman Grammar of Spoken and Written English*. Harlow: Longman.

Culpeper, J. (2012). *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Heuser, R. and Le-Khac, L. (2011). Learning to Read Data: Bringing out the Humanistic in the Digital Humanities. *Victorian Studies*, **54**(1): 79-86.

Ludwig, R. and Nault, C. (1989). *The Annals of American Literature, 1602-1983*. Oxford: Oxford University Press.

Mahlberg, M. (2012). *Corpus Stylistics and Dickens' Fiction*. New York: Routledge.

Underwood, T. and Sellers, J. (2012). The Emergence of Literary Diction. *Journal of Digital Humanities*, **1**(2).

Underwood, T. (2013). Wordcounts are amazing. *The Stone and the Shell*, 20 February. Retreived from: <http://tedunderwood.com/2013/02/20/wordcounts-are-amazing/>.

Notes

[1] In stylometry, there has been lively debate about the use of n-grams for authorship attribution. Our study differs in focusing on very common n-grams (i.e., on lexical bundles) rather than the rare n-grams studied by Vickers (2011), and it does so not for the purposes of authorship attribution, but rather to identify how fiction's most common lexical bundles may have evolved over time, and may have done so differently in narration and in dialogue.

[2] Jonathan Culpeper's 2012 book is another recent study that uses bundles to study literary texts, but Culpeper's object of inquiry is the early modern English dialogue, not the novel.

[3] Departing from Biber et al., we decided to tokenize contractions as separate words (with "won't" becoming "wo" + "n't"), in order to place bundles involving contractions in direct comparison with their uncontracted equivalents ("i wo n't" and "i will not" being both tri-gram bundles).

[4] Biber and Conrad (1999) argue, and we agree, that lexical bundles often serve as mechanisms for bridging syntactic and semantic units. Accordingly, we allow lexical bundles to cross punctuation as well as clause and phrase boundaries.

[5] For example, in the 1835 novel by William Gilmore Simms, *The Partisan: A Tale of the Revolution*: "but there was a reckless audacity in his replies to the friendly suggestions of the landlord, which half-frightened the latter personage out of his wits."

Performance, the Document, and the Digital: the Case of Lynn Hershman Leeson's 'Robertas'

Gabriella Giannachi
g.giannachi@exeter.ac.uk
University of Exeter, GB

Performance at Tate, an Arts and Humanities Research Council funded project which run between 2014-6, set out to trace the history of performance at Tate from the 1960s to today by investigating practices of collection, display, documentation and exhibition in the museum. At the heart of the project was the desire to conduct a wider re-evaluation not only of the place of performance in the museum, but also of the specific role played by documentation, including digital documentation, as well as the documentation of digital works, within collections, archives and displays. Here, I explore what the introduction of the digital has meant in this particular field by conducting a close examination of Lynn Hershman Leeson's *Roberta Breitmore* (1972-8), in which the artist created a fictional persona and interpreted its role for a period of six years using surveillance technology to capture various moments in her life. I will also discuss her subsequent works *CyberRoberta* (1995-8) and *Life to the Second Power* (2007) in which the character of Roberta was re-invented across different media. Focussing on the different types of documentations that these works generated, including photos, drawings, a cartoon, a film, a second-life re-enactment, postcards, among others, I then establish a best practice framework towards their curation and preservation that will be applicable more broadly for digital art practice.

To establish the role played by documentation in this context, Performance at Tate aimed to move beyond existing debates on the ontology of the relationship between performance and documentation. These debates may be traced back to the 1970s when, writing on performance-based work, the art historian Douglas Crimp asserted that 'you had to be there', implying that performance needed the presence of the spectator to be activated and often required 'that registration of presence as a means toward establishing meaning' (1979: 77). This approach underpins the performance studies scholar Peggy Phelan's well known assertion that 'performance's only life is in the present' and that performance 'cannot be saved, recorded, documented, or otherwise participate in the circulation of representations *of* representations: once it does so it becomes something other than performance' (1993: 146). A different position was adopted by media studies scholar Philip Auslander who in his identification of different types of performance (and documentation) counter-pointed that 'documentation does not simply generate image/statements that describe an autonomous performance' and states that it can produce 'an event as a performance' (2006: 5).

Instead, Performance at Tate aimed to build on approaches initiated by Amelia Jones (1997) and, subsequently, Barbara Clausen, who challenged the positioning of the document as secondary to performance, as well as the positioning of the document as equivalent to performance suggesting that performance should not be seen as beginning with or ending with the 'authentic experience', or live moment, but rather that it should be seen as 'an ongoing process of an interdependent relationship between event, medialization, and reception' (2005: 7). In other words, performance, in the course of its transcriptions, is subject to significant shifts caused by the constantly altering reception of its documents over time. Performance documents should therefore be considered, utilising Suzanne Briet's term from 1951, as an ecology of inter-documents, comprising primary documents, created at the time of an event, secondary documents, created from the initial documents, and auxiliary documents, created by a juxtaposition of documents. Rather than delivering remains of an isolated event, the document, for Briet, forms part of a matrix or network of signs. So, she noted, 'through the juxtaposition, selection, and the comparison of documents, and the production of auxiliary documents', the content of documentation becomes 'inter-documentary'.

Performance and documentation have always been somewhat inter-dependent. So, for example, art historian and critic Barbara Rose pointed out the significance of Hans Namuth's famous photographs of Jackson Pollock's work as *documents* of his practice that radically affected any subsequent perception of his paintings (1979: 12). Likewise, it was performance studies scholar Philip Auslander who noted that Harry Shunk's photographs of Yves Klein's *Leap into the Void* (1960), a photomontage, in fact constitute the work itself (2006). And it was Paul Schimmel who noted how Chris Buren's actions, such as in *Shoot* (1071), in which the performer asked his assistant to shoot him in his left arm, were 'distinguished by their ability to be captured by a single photographic image and described in a brief paragraph' (1998:97) almost as if to imply that the performance was designed so as to work for the photo. Most of these works nowadays exist primarily as documents. One such work is Hershman Leeson's *Roberta Breitmore* (1972-8) which comprises of a series of documents charting Roberta's internal (i.e. a list of cosmetics for her make-up) and external transformations (i.e. a movement chart), testifying also to her social existence (i.e., she placed an advert, and underwent a psychiatric evaluation) and financial existence (i.e. she owned a checkbook). Nearly twenty years after Hershman Leeson exorcised the character of Roberta at the Palazzo dei Diamanti in Ferrara in 1978, Roberta was re-invented as *CyberRoberta* (1995-8), a tele-robotic doll who was dressed identically to Roberta, and whose fictional persona was, as in Hershman Leeson's words, 'designed as an updated Roberta' who navigate the internet, and was described as a 'cyberbeing' (1996: 336). Roberta also appeared as a bot in the Second Life remake of an early work by Hershman Leeson, *The Dante Hotel,* called *Life to the Second Power* (2007-) , which turned parts of the Hershman Leeson archive at Stanford Libraries into a dynamic mixed reality experience where visitors could explore digital reproductions of fragments of the original archive of *The Dante Hotel* under Roberta's guidance in Second Life (Roberta had started her existence when she arrived in San Francisco on board of a Greyhound bus and checked herself in at the Hotel Dante).

The first work, *Roberta Breitmore,* consists of documents which are now preserved as the artwork in public and private collections (MOMA, SFMOMA, Tate, Walker Art Center, as well as the Hess collection, to name a few). Each museum also has a documentation of the work, which usually consists of gallery, curatorial, and preservation records. These may disclose significant information about how the artist wishes the work to be installed, for example, or about the work's preservation strategy. *CyberRoberta*, which is in the Hess collection, does not consist of documents, though, unlike in the case of *Roberta Breitmore*, some documents were produced by users and are available on social media. These documents are not works, though it is only by seeing them that, if we have not experienced the work ourselves, we can understand how the work operated. No museum, to my knowledge, has been preserving these user-generated documents. Finally, *Life to the Second Power* is available on Second Life and a set of photographic documents were collected by Exeter and Stanford Universities at the time the work was shown. Again, a number of visitors generated photographs (both in first and second life) but these were not collected. The work was shown as The Montreal Museum of Fine Arts and at SFMOMA, so it is likely that these museums have kept a documentation of the work, though most museums only do so when the work entered the collection. In short, in the case of *Roberta Breitmore,* the artist created the documents that are now known as the work. In the cases of *CyberRoberta* and *Life to the Second Power,* most documents showing the work in use were produced by users or viewers, and none of them are systematically preserved. In the case of *Life to the Second Power,* the work is hosted by a commercial platform and may cease to exist once the platform becomes obsolete or is terminated by its owners.

I conclude by suggesting that Museums should draw from performance studies and digital humanities and create records documenting the experiences of these works, noting also how these have changed over the years. I also look into the challenges paused by their preservation, particularly in terms of born digital works. Finally, I show that by capturing this knowledge, Museums will not only preserve important historical information about the exhibition and reception of these works, but also create, to use Briet's term, an inter-documentary ecology comprising 'live' performance (whether by the artist or the user), documents (created by the artist, the museum or the user) and the digital (showing the web and social media life of a work in different formats).

Bibliography

Auslander, P. (2006). The Performativity of Performance Documentation. *PAJ: A Journal of Performance and Art* 84, **28**(3): 1-10.

Briet, S. (2006). [1951]. *What is Documentation?* Tr. E. Day, L. Martinet, H. G.B. Anghelescu. Lanham, MD: Scarecrow Press.

Clausen, B. (2005). The (Re)presentation of Performance Art. *After the Act*, Museum Moderner Kunst Stiftung Ludwig Wien, pp. 7-20.

Crimp, D. (1979). Pictures. *October,* Spring 1979, pp. 75-88.

Hershman Leeson, L. (1996). ed., *Clicking In: Hot Links to a Digital Culture.* Seattle WA: Bay Press.

Jones, A. (1997). 'Presence' in Absentia: Experiencing Performance as Documentation. *Art Journal,* **56**(4): 11-8.

Phelan, P. (1993). *Unmarked: The Politics of Performance*. New York and London: Routledge.

Rose, B. (1979). Hans Namuth's Photograph and the Jackson Pollock Myth: Part One: Media Impact and the Failure of Criticism, *Arts Magazine,* **53**(7).

Schimmel, P. (1998). (Ed), *Out of Actions: between performance and the object 1949-1979*, London: Thames and Hudson.

RICardo Project : Exploring 19th Century International Trade

Paul Girard
paul.girard@sciencespo.fr
Sciences Po médialab, Paris, France

Béatrice Dedinger
beatrice.dedinger@sciencespo.fr
Sciences Po Centre d'histoire, Paris, France

Donato Ricci
donato.ricci@sciencespo.fr
Sciences Po médialab, Paris, France

Benjamin Ooghe-Tabanou
benjamin.ooghe@sciencespo.fr
Sciences Po médialab, Paris, France

Mathieu Jacomy
mathieu.jacomy@sciencespo.fr
Sciences Po médialab, Paris, France

Guillaume Plique
guillaume.plique@sciencespo.fr
Sciences Po médialab, Paris, France

Grégory Tible
gregory.tible@sciencespo.fr
Sciences Po médialab, Paris, France

RICardo (*Research on International Commerce*) is a database gathering bilateral flows of international commerce extracted from a large number of historical statistical sources during the 1787-1938 period. The foundational principles of this database are described in a working paper (Dedinger and Girard, 2015) submitted to the Historical Methods journal. This database begun to be developed by researchers in Economy and History in 2004 with the goal of renewing research on the history of trade globalization. In 2013, an entire new direction was given to the RICardo project. Economists and historians worked together with data scientists and designers in order to build a completely innovative digital tool susceptible to be used for both teaching and research[1].

Exploring international trade during the 19th century

To this day, there is no equivalent digital resource to the RICardo database that focuses on such an ancient and lengthy time period. The only comparable tools are web applications from the IMF (data.imf.org), the WTO (stat.wto.org) and the United Nations (comtrade.un.org) that offer data and visualizations on commerce during the post-50's period. As we will show, the methodological problems posed by historical trade statistics (19 th-20 th centuries) are more complex. The first result of our work is a web application [2] available online:http://ricardo.medialab.sciences-po.fr. The database will be publicly released in 2017 upon the start of an international conference marking the two-hundredth anniversary of the publication of Ricardo's main work (Ricardo, 1817).

We will discuss the methodological choices made during the creation of this data exploration tool – what we call *datascape* (Latour et al., 2012) – and will quickly present the research and educational perspectives that the tool allows for, before concluding on the method of transdisciplinary work that was used.

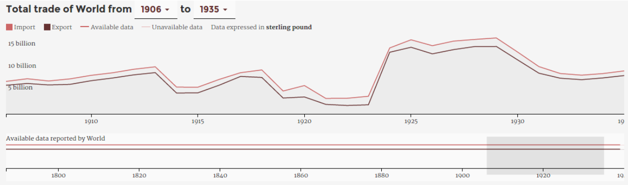
Representing data in their uncertainty

In RICardo, the basic informational unit is a trade flow (exportation or importation) between a *reporting* unit and a *partner* unit. In its present version, the database contains

267 000 flows. The *partners* are the commercial partners recorded in the annual trade statistics reports of the *reporting* countries. The large spatial and temporal coverage of the database raises certain problems. First off, the entities cover very heterogeneous realities: partners can be countries ("United Kingdom"), groups of countries ("United Kingdom & Ireland") or geographic areas ("British colonies"). Further, the availability of trade statistics before the end of the 19 th century is very problematic (Dedinger and Girard, 2015), which is translated into an absence, in the database, of a non-negligible (17 000) amount of flows throughout the observed period. The challenge we then had to take up was to aggregate heterogeneous data without crushing the corpus' complexity and accounting for missing data in the series. We attempt to resolve this thorny question by the exploratory analysis of the data (Tukey, 1977).

Three levels of exploration

The exploration interface developed offers three levels of entry from global to local: *world view, country view* and *bilateral view*. Each representation offers a specific point of view on the data within the database. The complexity increases as we move from the world view (*World*) to the national point of view (*bilateral*). In order to lead the exploration throughout these different levels, we have decided to construct a common structure to all three viewpoints.


Total trade of World from 1906 to 1935

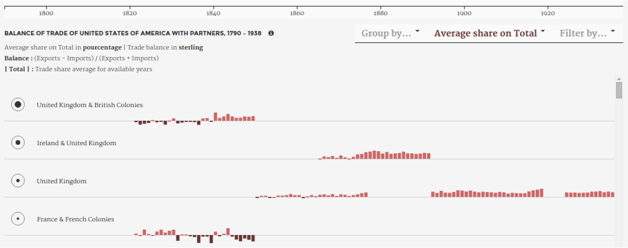
Data expressed in **sterling pound**

Figure 1: Temporal filtering

Each one opens unto a main curve representing total flows per year (representation of countries' total trade for the *world view*, of one country's total trade for the *country view*, or tradebetween two specific countries for the *bilateral view*) in addition to temporal limits. It is represented as a discontinuous line that is interrupted when there is no known value. In addition, these discontinuities appear also in an underlying temporal axis, indicating the absence of data in the form of a projection on the X-axis. This graphic object is also useful as an interface for temporal filtering: one can graphically select a sub-period to study more specifically (Becker and Cleveland, 1978). As footer for each view's page, a spreadsheet allows to navigate within the visualized source data and export it as spreadsheet format (CSV). This way, the user can continue the analysis in an external statistical tool. At the center of this common structure, each view offers more detailed exploration.

Studying a country through the prism of its partners

The country view allows users to center their analysis on a *reporting* country by representing annual trade balances with each *partner* under the form of histograms.


BALANCE OF TRADE OF UNITED STATES OF AMERICA WITH PARTNERS, 1790 - 1938

Average share on Total in **pourcentage** | Trade balance in **sterling**

Balance : (Exports − Imports) / (Exports + Imports)

| Total | : Trade share average for available years

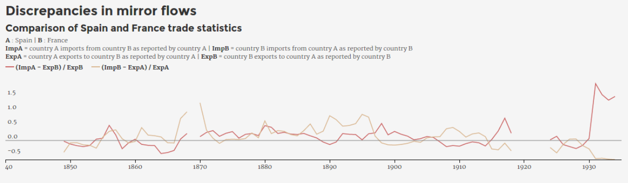
Group by... ▾ Average share on Total ▾ Filter by... ▾

Figure 2: Trade balances of partners

This use of *small multiples* (Tufte, 1990) facilitates comparisons between partners. The choice of detailing trade balances per year holds the fact that an aggregation on the entirety of the period would have introduced a bias by masking the year in which some data is missing. Furthermore, the ordering of the partners corresponds to the average annual share of each partner on the available years. This metric, represented as a circle at the start of the line, allows on the one hand to display the partners by decreasing importance, and on the other hand to let appear similar *partner* entities. Figure 2 therefore shows that "Ireland & United-Kingdom" and "United Kingdom" were the 2 nd and 3 rd partners of the United States during the 19 th century. They are two different entities but a similar partner in reality: the United Kingdom (the overlapping of years 1864-1876 is a consequence of the inclusion of two different sources).

The distorting mirror of bilateral trade flows

In the bilateral view, the tool offers to interrogate a pair of *reporting* countries by using a representation of their mirror flows. It's one of the great strength of the RICardo database: a same bilateral flow is declared as an export by one of the countries and as an import by the other, however these two recordings are rarely the same (Dedinger and Girard, 2015). Thanks to the calculation of an indicator (Dedinger, 2012), the bilateral view offers an immediate view of the deviation between mirror flows and its fluctuations over the selected period.


Discrepancies in mirror flows

Comparison of Spain and France trade statistics

A : Spain | **B** : France

ImpA = country A imports from country B as reported by country A | **ImpB** = country B imports from country A as reported by country B

ExpA = country A exports to country B as reported by country A | **ExpB** = country B exports to country A as reported by country B

Figure 3: The bilateral view

A tool for researchers and students in social sciences

The exploratory formatting of data brings powerful tools to help research on the history of economics. It's an excellent way to detect possible inconsistencies in data (Leclercq et al., 2013) in order to determine its cause. To achieve such a result, our method consisted in considering data visualizations as both results and media of research (Stefaner, 2010), in fostering collaboration between researchers, engineers and designers through workshops where each data visualization was discussed by articulating the methodological constraints of Economics and History, the ideas and principles of Design and technical workability.

Researchers have at their disposal an incomparable tool to deepen an analysis of trade statistics reliability (bilateral view), to compose monographs on the trade history of the world's countries (country view), to study the history of trade globalization since the beginning of the 19 th century (world view) and to experiment data-driven teaching methods.

Bibliography

Becker, R. A. and Cleveland, W. S. (1987). Brushing Scatterplots. *Technometrics*, **29**(2): 127 doi:10.2307/1269768.

Dedinger, B. (2012). The Franco-German trade puzzle: an analysis of the economic consequences of the Franco-Prussian war1: THE FRANCO-GERMAN TRADE PUZZLE. *The Economic History Review*, **65**(3): 1029–54 doi:10.1111/j.1468-0289.2011.00604.x.

Dedinger, B. and Girard, P. (2016). Visualizing trade globalization in the long run : the RICardo project. *Working Paper Submitted to Historical Methods*http://ricardo.medialab.Sciences-po.fr/Dedinger_Girard_RICardo_HistoricalMethods.pdf (accessed 26 February 2016).

Latour, B., et al. (2012). 'The whole is always smaller than its parts' - a digital test of Gabriel Tardes' monads. *The British Journal of Sociology*, **63**(4): 590–615 doi:10.1111/j.1468-4446.2012.01428.x.

Leclercq, C. and Girard, P. (2013). The Experiments in Art and Technology Datascape. *Collections électroniques de l'INHA. Actes de Colloques et Livres En Ligne de l'Institut National D'histoire de L'art*. INHA http://inha.revues.org/4926 (accessed 26 February 2016).

Ricardo, D. (1817). *On the Principles of Political Economy and Taxation*. London: John Murray.

Stefaner, M. (2010). Bootstrapping - use visualizations to create visualizations Paper presented at the VisualEyes / The Role of Design in Data, Information and Knowledge Visualization, Politecnico Milano, Italy http://www.slideshare.net/MoritzStefaner/bootstrapping-use-visualizations-to-create-visualizations (accessed 26 February 2016).

Tufte, E. (1990). *Envisioning Information*. Cheshire, CT, USA: Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

Notes

[1] A French version of this abstract is available at: http://ricardo.medialab.sciences-po.fr/Girardetal_RICardo_dh2016_fr.pdf

[2] the source code is available at http://github.com/medialab/ricardo

An Augmented Reality Mobile Application for Intergenerational Learning and Critical Connection: Experiencing The Histories of Flushing Meadows Park

Tamar Gordon
gordot@rpi.edu
Rensselaer Polytechnic Institute, United States of America

Lori Walters
lcwalter@ist.ucf.edu
Institute for Simulation and Training, University of Central Florida

Rob Michlowitz
rob.michlowitz@gmail.com
Institute for Simulation and Training, University of Central Florida; Institute for Simulation and Training, University of Central Florida

Augmented Reality (AR) superimposes geolocated digital text and images on a real-time view of the world. Museums, urban spaces and cultural heritage sites are using AR to virtually restore and enhance historical displays and environments on smartphones or tablets. We have created an AR application that combines dynamic, interactive content, including realistic 3D models, video, animation, game, and website portal, to immerse a diverse public in the history and culture of an important and well-trafficked setting: Flushing Meadows-Corona Park (FMCP), a 1255-acre urban park in Queens, NY. Littered with dramatic remnants of two World's Fairs, FMCP is steeped in a largely-forgotten history as the site of the 1939/40 and 1964/65 World's Fairs (NYWF), the first UN General Assembly, activities of European colonists, and early habitation by Algonquian speaking Native Americans. Additionally, the park served as the backdrop for the novel *The Great Gatsby* and as a site for an interesting natural history.

Our public venture partner, the Queens Museum, is situated within FMCP, in the most diverse immigrant neighborhood in New York State. In 2014 the population of Queens was composed of 49.1% Caucasian, 28% Latino, 25.8% Asian, and 20.8% African American. 47.8% of households are headed by individuals born outside the United States (http://quickfacts.census.gov/qfd/states/36/36081.html). By harnessing the multimodal possibilities of AR, we have engaged an ethnically, generationally and educationally diverse audience in the cooperative - and critical - exploration of history and culture.

The application will provide users two complementary experiences: an intergenerational learning game which enables children and adults to be partners while learning about the history of FMCP, and a guided tour that explores the two World's Fairs through comparison of their common, recurrent themes such as futurism, technological optimism, citizenship, progress, race and nationalism. Through these modalities, the application supplies a greatly enhanced destination experience to users of all ages and backgrounds, whether they are families from the surrounding neighborhoods, Queens Museum goers or tourists to New York City.

The project is informed by interdisciplinary humanities scholarship, including World's Fair history, anthropology, theme park studies, cognitive psychology, AR mobile design,and game design for STEM and cultural learning (http://srealserver.eecs.ucf.edu/chronoleap/).

I. Intergenerational Game

Based upon the activity Geocaching, an intergenerational learning game is under development to provide children a scaffolded way to learn about the history and heritage of FMCP. Geocaching is a scavenger hunt-style game, which sends individuals or teams searching an environment for hidden physical items. By making the caches virtual, there are no limitations on items and avoid impacting park operations. An adult led by the child must surmount the challenge of locating virtual geocaches, while exploring the past of FMCP. Teams will also drill virtual core samples in the park and examine them to find geocached items within.

Game play proceeds as users are provided clues and instructions guiding them to a physical locations. For example, a clue might read "Find the ancient column" (referring to the extant Roman Column of Jerash, presented to NYC by Jordan's King Hussein at the close of the 1964-65 NYWF) (Wingfield, 2011). Player's devices notify them when reaching the location and are then provided new GPS coordinates. The device enables them to find the location, as well as see the landscape of the park as it appeared in the past and find the virtual artifact. Artifacts are stored by the application for future examination. This process continues for a set of artifacts – each artifact being one piece of a puzzle. In geocaching tradition we provide a final reward at the end.

Another feature entails the collection of virtual pieces of artifacts from the core-sample's strata, which are eventually reassembled at the relevant location. Examples of artifacts which might be found and assembled by users include the Videophone displayed at the Bell Telephone Pavilion during the 1965/65 Fair, the first television displayed at the RCA Pavilion during 1939/40 Fair, and a Mastodon from the Pleistocene Era. The application can provide supplemental content (e.g. video, audio and images), questions that open up further pathways for thought, and fun quizzes that challenge a user's knowledge and memory.

As children have yet to develop a mastery of most subjects, they can often learn better when they are provided support from a familiar adult (Vygotsky, 1978). Such support, commonly known as scaffolding, is essential for them to integrate new information into their base of knowledge. This dovetails well with Lave and Wenger (1991), where learning is a communal, interactive activity where less experienced individuals learn from master practitioners. Our app provides an intergenerational pathway, where the adult can both support learning information presented and share associated information from their own life. A child will likely to engage with media prompts when with an adult (Takeuchi et al., 2011). Participatory mediation of learning, where adults and children collaborate, sharing a dialogue stimulated by the geocaching activity offers the potential for learning in such application (Clark, 2011). This could be extended to encompass the sharing of deep-rooted memories and experiences for the adult, while also enabling children to share their experiences.

II. The Deep Experience

This modality targets individuals who are interested in advanced content and interpretation for deeper, critical understanding of the 1939/40 and 1964/65 Fairs at Flushing Meadows, remnants of which form a dramatic backdrop. World's Fairs were typically experienced as spectacles of culture, industry, and technology. (Benedict,1981; Geppert et al., 2005; Greenhalgh, 2011; Corn and Horrigan, 1984; Garn et al., 2007; Rydell, 1984, 1993; Rydell and Gwinn, 1994; Rydell et al., 2000) However, they are also environments that shaped popular understanding of society's past, present and future, linked through enduring narratives of citizenship, progress, technological optimism, consumerism, race and nationalism (Hollengren et al., 2014; Gordon, 2005; Marchand ,1991,1995; Samuel, 2007; Tirella, 2014; Winner, 1991). This mode enables users to critically engage with the changing social themes and narratives that tie the Fairs together by sending users to cognate, virtual pavilions.

Designed as a curated tour starting at the Queens Museum, visitors are guided through the park viewing models of Fair pavilions, and learning about common themes through archival images, audio and documentary footage, both within the application itself and on the associated website. Examples include:

Landscapes of the Future

The connected themes of progress, technology, commerce and utopia – as expressions of "the future in the present" – have always dominated in World's Fairs (cf Land, 2011). Visitors are directed to the interstate highway utopia of Futurama at the General Motors pavilion, designed by Norman Bel Geddes (Marchand, 1991; 1995), and also to Futurama II, its updated version in the later Fair, in which "General Motors set out to reveal how technology would conquer the harshest environments for the betterment of humanity" (Walters, 2014:467). Visitors will be able to view archival footage of both exhibits, read contemporary accounts and see original designs.

Protests at the Fair

National and global conditions did not reflect the optimistic themes of the Fairs, nor did Fairs go uncontested by groups who were excluded from self-representation and employment. The 1964/65 NYWF was designed as a virtual bubble away from the emergent civil rights, free speech and anti-war movements. Visitors are invited to explore the CORE (Congress of Racial Equality) protests of 1964 (Peneda, 2014; Jacoby, 1998; Tirella, 2014; http://www.democracynow.org/2014/4/25/protesting_the_1964_world_s_fair)

and to see them come alive at the site of the Unisphere (still extant), through digital overlay of archival footage (see for example http://timetraveler.berlin). Visitors will also access additional material such as pamphlets, flyers and photographs, and hear audio of speeches and meetings.

Racial Representation: two Africas

Both Fairs contained pavilions that represented "Africa" as a project of defining, organizing and displaying people and culture. (Lorini, 1999; Rydell, 1999). Our comparison of these exhibitions highlights the increased agency of the emergent African Republics of 1964/65 to represent themselves as modern members of the family of nations with professional dance performances, art exhibit, displays of industry (cf Benedict, 1993; Lukas, 2007). In contrast, the Great Britain pavilion of 1939/40 contained "natural history" dioramas of traditional scenes in the life of its colonized subjects (http://www.1939nyworldsfair.com/worlds_fair/wf_tour/hall_of_nations/great_britain, https://archive.org/stream/ldpd_11290477_000#page/n1/mode/2up). The static diorama was a key visual strategy that distinguished colonized Africans from its white Commonwealth citizens, also on display. Visitors will be able to compare the 1964/65 professional dance performances with those that took place at the 1939/49 Midway where they were framed as exotic, marginal entertainments.

The interpretive experience provides an associated website that contains contemporary accounts, academic essays, archival media, and exhibition content. The website will also provide an opportunity to upload personal oral histories of the Fairs (cf Anderson, 2003) that will prompt them to critically reflect on the relationship of the historic Fairs and contemporary life.

Bibliography

Anderson, D. (2003). Visitors Long-term Memories of World Exhibitions. *Curator: The Museum Journal,* **46**(4): 401-20.

Benedict, B. (1991). International Exhibitions and National Identity. *Anthropology Today,* **7**(3): 5-9.

Benedict, B. (1981). *The Anthropology of Worlds Fairs.* Scolar Press.

Clark, L. (2011). Parental Mediation Theory for the Digital Age. *Communication Theory*, **21**(4): 323–43. doi:10.1111/j.1468-2885.2011.01391.x.

Clasen, W. (1968). *Expositions, Exhibits, Industrial and Trade Fairs.* New York: Praeger.

Corn, J. J. and Horrigan, B. (1984). *Yesterday's Tomorrows: Past Visions of the American Future.* Washington DC: Smithsonian Institution Press.

Corn, J.J. ed. (1986). *Imagining Tomorrow: History, Technology, and the American Future.* Cambridge, MA: MIT Press.

Eco, U. (1986). A Theory of Expositions. In *Travels in Hyperreality*. San Diego: Harcourt Brace Jovanovich, pp. 291-307.

Garn, A., Antonelli, et al. (Eds.) (2007). *Exit to Tomorrow: World's Fair Architecture, Design, Fashion 1933-2005*. New York: Universe Publishing.

Geppert, A. C. T., Coffey, J. and Lau, T. (2016). International Exhibitions, Expositions Universelles and World's Fairs, 1851-2005, *A Bibliography*. http://www.csufresno.edu/library/subjectresources/specialcollections/worldfairs/ExpoBibliography3e d.pdf (accessed 2 January 2016).

Gordon, T. (2005). *Global Villages: The Globalization of Ethnic Display.* Tourist Gaze Productions, 58 mins.

Greenhalgh, P. (2011). *Fair World: A History of World's Fairs and Expositions from London to Shanghai 1851-2010*. London: Papadakis Press.

Herbst, I., et al. (2008). *TimeWarp: Interactive Time Travel with a Mobile Mixed Reality Game*. Paper presented at MobileHCI 2008, Amsterdam.

Hollengreen, LH., et al. (Eds.) (2014) *Meet Me at The Fair: A World's Fair Reader*. Pittsburgh, PA: Carnegie Mellon ETC Press.

Land, N. (2011). *Neomodernity*. Urban Future 22.

Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press.

Lorini, A. (1999). International Expositions in Chicago and Atlanta: Rituals of Progress and Reconciliation. In Lorini, A. *Rituals of Race: American Public Culture and the Search for Racial Democracy*. Charlottesville: University Press of Virginia, pp. 33-75.

Lukas, S. (2012). *The Immersive Worlds Handbook: Designing Theme Parks and Consumer Spaces.* New York, NY: Focal Press.

Lukas, S. ed., (2007). *The Themed Space: Locating Culture, Nation, and Self.* Lanham, MD: Lexington Books.

Marchand, R. (1991). Corporate Imagery and Popular Education: World's Fairs and Expositions in the United States, 1893-1940. In David E. Nye and Carl Pedersen (Eds.), *Consumption and American Culture.* Amsterdam VU University Press, pp. 18-33.

Marchand, R. (1995). The Designers Go to the Fair, I: Walter Dorwin Teague and the Professionalization of Corporate Industrial Exhibits, 1933-1940; The Designers Go to the Fair, II: Norman Bel Geddes, the General Motors "Futurama," and the Visit-to-the-Factory Transformed. In Doordan, Dennis (ed.) *Design History: An Anthology*. Cambridge, MA: MIT Press, pp. 89-121.

Peneda, E. (2014). Present Tense, Future Perfect: Protest & Progress at the 1964 World's Fair. *Appendix* **2**(3): 118-26. http://theappendix.net/issues/2014/7/present-tense-future-perfect-protest-and-progress-at-the-1964-worlds-fair. (accessed 1 February 2015)

Rydell, R. W. (1984). *All the World's a Fair: Visions of Empire at American International Expositions, 1876-1916*. Chicago, IL: University of Chicago Press.

Rydell, R. W. (1993). *World of Fairs: The Century-of-Progress Expositions.* Chicago, IL: University of Chicago Press.

Rydell, R. W. and Gwinn, N. E. (Eds.) (1994). *Fair Representations: World's Fairs and the Modern World.* Amsterdam: VU University Press.

Rydell, R. W. (1999). "Darkest Africa": African Shows at America's World's Fairs, 1893-1940. In Lindfors, B. (ed), *Africans on Stage: Studies in Ethnological Show Business.* Bloomington: Indiana University Press/Cape Town, pp. 135-55.

Rydell, R. W., Findling, J. E. and Pelle, K. D. (2000). *Fair America: World's Fairs in the United States.* Washington DC: Smithsonian Institution Press.

Samuel, L. (2007). *The End of the Innocence: The 1964-65 NY World's Fair*. Syracuse, NY: Syracuse University Press.

Sorensen, C. (1989). Theme Parks and Time Machines. In Vergo, P. (Ed). *The New Museology*. London: Reaktion Books, pp. 60-73.

Stanley, N. (1998). *Being Ourselves for You: The Global Display of Cultures*. Middlesex University Press.

Takeuchi, L., et al. (2011). "The New Coviewing." http://www.joanganzcooneycenter.org/wp-content/uploads/2014/03/JGC_CoViewing_iPadPrint.pdf. (accessed 3 February 2016).

Tirella, J. (2014). *Tomorrow-Land: The 1964-65 World's Fair and the Transformation of America.* Guilford, CT: Lyons Press.

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Walters, L (2014). The Japan Pavilion at the 1964/65 New York World's Fair: A Vision of the Near Future. In Hollengreen, LH., Pearce, C., Rouse, R., Schweizer, B., (Eds.). *Meet Me at The Fair: A World's Fair Reader*. Pittsburgh, PA: Carnegie Mellon ETC Press, pp. 467-72.

Wingfield (2011). *Whispering Column of Jerash*. New York Public Library Blog. http://www.nypl.org/blog/2011/10/24/whispering-column-jerash (accessed 27 January 2016)

Winner, L. (1991). An Alternative Worlds Fair Could Playfully Debunk Myths about Technological Progress. *Technology Review,* 94.

Yovcheva, Z., Buhalis, D. and Gatzidis, C. (2013). Engineering Augmented Tourism Experiences. In L. Cantoni, and Z. Xiang (Eds.), *Information and Communication Technologies in Tourism 2013*, pp. 24-36. Heidelberg: Springer.

Bringing Migration Data Into Context Using Digital Computational Methods

Ronald Haentjens Dekker
ronald.dekker@huygens.knaw.nl
Huygens ING, Netherlands, The

Rik Hoekstra
rik.hoekstra@huygens.knaw.nl
Huygens ING, Netherlands, The

Marijke van Faassen
marijke.van.faassen@huygens.knaw.nl
Huygens ING, Netherlands, The

Introduction

Preparing sources for historical research usually requires making many heterogeneous collections digitally accessible and linking them to compose a multi-faceted and multi-layered resource that supports both distant reading and close reading forms of analysis. In the lifecycle of historical information – a model introduced in 2004 - the Dutch DH-experts Boonstra, Breure and Doorn emphasize three points that should be kept in mind by e-science experts and researchers alike to keep historical information systems alive and useful: durability, usability and modeling (Boonstra et al., 2006: 22).

Timbuctoo, developed at the Huygens ING, offers a system that makes it possible to model and store heterogeneous data but also incrementally enrich and link the data. Furthermore, it also offers facilities to document the provenance of all data as well as all steps in data editing, extraction and linking data. These features are vital for historical research in which researchers need to be able to exert 'source criticism' and go back to the original source or data at all times (Ockeloen et al., 2013).

We will demonstrate the solution Timbuctoo offers with the Migrant, Mobilities and Connection project as a use case, because of its complex and multiple links to datasets from a myriad of cultural heritage institutions (archives, libraries and museums) on several levels.

The main focus of the Migrant project is on the life courses of the migrants. Starting from the limited core data from a connection between (digitized) Dutch emigrant cards and Australian immigration files, the life courses will be elaborated using in depth-analysis of these and other collections. It is important to note that for the purpose of the Migrant project that life courses not only comprise dates and birth, marriage, migration, education and employment but also extend to the interactions of migrants with all sorts of institutions in the Netherlands and Australia and their representatives. The database therefore enables us to analyse and compare the evolution of a multitude of social networks (Arthur et al. (submitted); Van Faassen, 2014a, 2014b).

Migration files as a multi-layer resource

The Migrant, Mobilities and Connection project focuses on the Dutch-Australian post World-War II migration from the Netherlands to Australia. Like all migrants these 180,000 people have left many traces in different cultural heritage collections ranging from (supra) government archives to the photo and memorabilia collections of the migrant families themselves and anything in between. These collections are dispersed over different countries. A lot of the collections are available in a digital form or will be digitized in the future, but like all historical collections they contain partially structured information and partially unstructured information that needs to be made accessible for further analysis. In elaborating the data we will use a variety of methods ranging from computer assisted data extraction and linking of a large collection of life events to hand editing of handwritten registration cards and personal migrant files. From an analytic perspective all these collections and edited data can be seen as different layers that need to be accommodated by the data store in which they are kept (Hoekstra and Nijenhuis, 2012).

Timbuctoo

Timbuctoo is a data repository system aimed at humanities research with the aim of linking together datasets containing structured information concerning people, places and organisations without actually merging them to facilitate scientific analysis and discussion.

To accomplish this it defines a number of primitive types that describe entities that all researchers agree on, such as the afore mentioned persons, places, organisations as well as works, languages, concepts and events. Each research project that makes use of the repository can extend the primitive types with extra fields. On top of that the repository has the ability to store multiple viewpoints on the same entity. In this way, researchers become aware of the different or sometimes even conflicting assumptions about entities, fueling scholarly debates in a conceptual way. Timbuctoo also support versioning and provenance. To make it clear on which information the results are based every change made to the data should have information who made the change, when the change was made and for what reason the change was made. The user interface, analysis and visualization are completely separated from the storage of the data. All services are coupled using REST (Fielding, 2000) APIs. The software is freely available under an open source license (GPL 3) and is published on Github.

Information extraction

During the project data will be added, edited and analysed continuously. As indicated above, at the beginning of the project the data consists of migration information contained in cards, files and information of governance agencies. Apart from the core information already available in a simple database, the cards and files contain much more information that must be digitized to be able to use it for analysis. In the course of the project, a lot of other materials from archives, libraries and other collections from different cultural heritage institutions will be added to the Timbuctoo database. Some of the information will be structured, but most is contained in typed or handwritten files and in images. The aim of information extraction is to extract structured information from unstructured information.

For the elaboration of this wealth of materials we will use an eclectic mix of editing and information extracting methods. Previously, hand editing was the only option for these types of materials, but in light of the amount of material that will be collected, all computer assisted information extraction that is possible will contribute to the database and help analysis.

To automate this process the data needs to be stored in such a way that a context can be build. The computer can search for patterns and suggest links to data already present in the network or calculate statistics to point out interesting or unusual things in the dataset. To begin with, an algorithm needs to be made to link the Dutch and Australian records together. Note that the system should not actually merge the records. Data about persons can be linked together based on (for example) familyname, year and place of birth and indeed all other types of structured information available such as migration date, migration scheme, ship with which they travelled or still very different data depending on the source and the context.

Other examples of computer assisted data extraction include the recognition of certain keywords in the facsimile or transcriptions. Researchers can add or edit information manually through the user interface, either manually or using algorithms. Automatic information extraction should suggest relations between different records or entities, but never actually enforce those changes on the researcher.

We started out to build Timbuctoo with the problem of the large variety of heterogeneous data sources that our institute produced in the course of a hundred years of classical and some twenty-five years of digital source editing and publishing. The use case of the Migrant, Mobilities and Connection project, with data about thousands of personal migrant stories scattered all over the world and its myriad of policy files on national and supranational levels recorded in different datasets, demonstrates the different features of Timbuctoo.

First, Timbuctoo is used as a repository where researchers as well as automatic tools such as parsers store all the heterogeneous data and the relations between it. Since all the data is versioned and provenance information is recorded, it is always clear where the data originates from. Second, Timbuctoo serves as a data source for researchers and parsers. Named entity recognition tools can use all the available names of places, organisations and persons as training data. Researchers can do queries on certain properties and do statistical analysis on the results to either find outliers or confirm or refute a hypothesis on a larger scale in more varied ways than previously possible. Finally, being a graph database, Timbuctoo is a research tool that enables researchers to infer indirect relations from the numerous direct relations in the repository. This makes it possible to perform complex queries and conduct network analysis and visualization. These three features combined enable researchers to discover unexpected phenomena that can lead to new research and methodological questions.

Bibliography

Arthur, P., et al. (2015 submitted). Migrating People, Migrating Data. *Digital approaches to migrant heritage (manuscript).*

Boonstra, O., Breure, L. and Doorn, P. (2006). *Past, present and future of historical information science*. Amsterdam: DANS.

Faassen, M. van (2014a). *Polder en emigratie. Het Nederlandse emigratiebestel in internationaal perspectief 1945-1967*. The Hague: Huygens ING.

Faassen, M. van (2014b). *Naoorlogs emigratiebeleid in Nederland 1945-1967*, The Hague: Huygens ING (http://resources.huygens.knaw.nl/emigratie).

Fielding, R. T. (2000) Architectural Styles and the Design of Network-based Software Architectures, *Doctoral Dissertation*. University of California.

Hoekstra, R. and Nijenhuis, I.J.A. (2012). Enhanced Access for a Paper World, conference paper for *Editing Fundamentals: Historical and Literary Paradigms in Source Editing. Ninth Annual Conference of the European Society for Textual Scholarship*. Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam.

Ockeloen, N., et al. (2013). BiographyNet: Managing Provenance at multiple levels and from different perspectives. In: *Proceedings of the Workshop on Linked Science (LiSC) at ISWC 2013*, Sydney, Australia, October 2013.

Cross-Institutional Music Document Search

Andrew Hankinson
andrew.hankinson@mail.mcgill.ca
University of Oxford, United Kingdom

Reiner Krämer
reiner.kramer@mcgill.ca
McGill University, Montreal, Canada

Julie Cumming
julie.cumming@mcgill.ca
McGill University, Montreal, Canada

Ichiro Fujinaga
ich@music.mcgill.ca
McGill University, Montreal, Canada

The Single Interface for Music Score Searching and Analysis (SIMSSA) project is building tools and best-practices for performing large-scale document image recognition, analysis, and search on music documents. In this paper, we will describe a novel technique for providing cross-institutional music document image search, allowing for the creation of a search engine for the contents of the world's music collections through a single search interface.

This paper will describe our methodology for building large-scale search systems that operate across institutions. We will describe the optical music recognition (OMR) process, which, like optical character recognition (OCR) for text, extracts symbolic representations from document images and places them in a structural representation for further processing. We will then describe our techniques for music analysis, extracting patterns for indexing the musical contents of these images into searchable representations. Finally, we present our efforts at building a system that will allow users to search musical documents from many institutions and retrieve the digitized document image.

Challenges

Perhaps the most significant challenge to building a global document image search system is how to retrieve, store, process, and serve document page images. These images have been produced through mass digitization efforts by individual institutions. Aggregating document images to provide cross-institutional document search has traditionally been provided through centralized efforts, where a single organization collects digital images and performs document recognition (i.e., OCR) on them.

While this approach provides a central tool for users to search and retrieve document images, it has several disadvantages. It often requires significant storage capabilities, as the central organization must store and manage all the images from its partner institutions. There are logistical challenges, integrating cataloguing data from multiple document collections and maintaining up-to-date information and error-fixes from the partner organizations. There are also legal implications over the ownership and copyrights of document image surrogates, even on out-of-copyright documents (Allan, 2007). This typically requires negotiations and embargoes on who can access certain types of content which differ across partner institutions, and which must be applied at the central organization level (HathiTrust, 2015).

These technical, legal, and logistical challenges may be mitigated if the partner organizations were able to host and control access to their images directly. Until recently, however, direct access to the document images hosted by an institution was difficult as it required interacting with a wide variety of digital repository software, each with their own particular ways of storing and serving images. There were no standardized methods to specify how a document image could be accessed directly in these repository systems.

Interoperable Image Collections

The International Image Interoperability Framework (IIIF) (Snydman et al., 2015) is a new initiative that attempts to standardize methods for retrieving digital images from an institution's digital image collection. The IIIF specifies two mechanisms for this, the Image API and the Presentation API. The Image API sets out a standard URI-based request format to which IIIF-compatible systems must conform. Using this URI format one may specify the size, region, rotation, quality, and format of the requested image, as well as basic information about the image. The Presentation API is used to describe structural and presentation information about an image, or a sequence of images. The Presentation API is structured using JavaScript Object Notation (JSON), which may then be parsed by other software, and within which pointers to images using the Image API are stored.

To give an illustration, a digitized book may be represented as a IIIF Presentation API manifest file. Each page image within the book would be retrievable by a URI to the page image stored on a remote server. To view the book, the manifest would be loaded into a IIIF-compatible image viewer, which would then fetch and load each of the document images and present them in sequence.

The typical use case for a IIIF manifest is for the purposes of retrieving and viewing document page images. However, we are proporsing a novel application of IIIF as a standard interface to perform document image recognition tasks on digital collections from many different institutions.

Distributed Document Image Recognition

We are building a web-based document recognition system, named Rodan, for performing optical music recognition (OMR) on large quantities of page images (Hankinson, 2014). Rodan is a workflow system, where different image processing, shape recognition, and document processing tools can be chained together to produce the sequence of discrete steps through which an image must proceed to extract the symbolic music representation of the content. Crucially, the exact cartesian positions of every musical symbol on the image are stored, providing a way to correlate the musical content with its physical position on the page image (Hankinson et al., 2012).

By providing Rodan with a IIIF Presentation API manifest, the document page images may be downloaded and the symbolic music notation extracted. However, rather than storing the image, we store just the IIIF Image API-formatted URI back to the original image. This allows us to discard the downloaded image file but point back to the image hosted by the originating institution. This approach eliminates the need to store and serve the images on our own systems, while still providing content-level access to document images hosted in different institutional repositories.

Music Analysis

Within music notation there are several levels of representation. The most basic level is that of the symbol–the graphical element printed on the page. Structures such as melodies, phrases, and cadential patterns are built from these symbols, and exist in multiple overlapping hierarchies; a phrase might contain a number of cadential patterns. A music search system must understand the different levels and structures in a musical work, beyond simply understanding the individual notes, as these structures may form structural objects that a user may wish to retrieve. Within the SIMSSA project we are developing tools and techniques for extracting patterns from symbolic music representations using the Music Encoding Initiative (MEI) and other structured music representations (Schubert and Cumming, 2015; Sigler et al, 2015).

The Vertical Interval Successions (VIS) (Antiilla and Cumming, 2014) tool we are developing provides a platform on which pattern analysis and extraction methods may be built. Like the document recognition process, VIS operates on the principle that computational music analysis is a sequence of tasks, where each task is responsible for extracting specific types of information that may then be passed on to subsequent tasks. In this way, the underlying symbolic representation of music notation may be used to build higher-level representations, which may then be sent to an indexing service for use in query and retrieval tasks.

Cross-Institution Indexing and Retrieval

After analysis, the symbolic representations and the structures of the music documents are indexed for retrieval in a search engine. The IIIF Image URI associated with the page image, stored in the document recognition stage and carried along in the analysis stage, provides the mechanism through which the page image may be retrieved from host institutions in response to a query on the symbolic music contents. Through this system, musical full-text (or "full-music") search can be performed on document images hosted and served from IIIF-compatible digital collections. Additionally, metadata and cataloguing data may be embedded in the IIIF Manifest, or linked to other machine-readable representations. This data may also be centrally indexed, allowing users to retrieve documents across institutions with useful textual searches such as titles, composers, or dates.

Impact and Future Work

With cross-institutional music document image search, institutions may make their collections available to a broader audience without the need to host their images with a third-party service. With IIIF-compatible image and manifest services, the barriers to entry for these institutions to provide these capabilities is relatively low; the metadata and images are already part of their digital infrastructure. Furthermore, by serving the images and metadata directly from their own infrastructure, institutions can track collection usage patterns through their own server analytics.

More general applications of this methodology will have significant impacts on libraries, archives, and other institutions' document image collections. By providing machine-readable access to document images directly, third-party services for document analysis, including distributed optical character recognition (OCR) may be built and deployed. This will have implications on large-scale computational re-use of digital resources, and will open up document image collections to distributed analysis by a global audience.

We are currently in the process of building a prototype system that incorporates all elements of the process described in this paper. Our existing tools, Rodan and VIS, are currently being used in research and production, with a third system in development that will provide a platform for developing search and retrieval tools.

One of our biggest unanswered questions concerns the human side of retrieval. With large quantities of recognized musical content, what sorts of tools and interfaces will people use to query the symbolic content of music documents? How will they conceptualize their symbolic music information needs, and what types of interfaces will they use to express these needs to a search system? What types of musical patterns will we need to extract from our musical documents to provide a useful symbolic search system? All of these questions we hope to investigate with a completed system.

Bibliography

Allan, R. (2007). After Bridgeman: Copyright, museums, and public domain works of art. *University of Pennsylvania Law Review*, **155**(4): 961–89.

Antilla, C. and Cumming, J. (2014). The VIS Framework: Analyzing counterpoint in large datasets. In *Proceedings of the Conference of the International Society for Music Information Retrieval*. Taipei, pp. 71–6.

Hankinson, A. (2014). "Optical music recognition infrastructure for large-scale music document analysis." PhD diss., Schulich School of Music, McGill University.

Hankinson, A., et al. (2012). Digital Document Image Retrieval Using Optical Music Recognition. In *Proceedings of the Conference of the International Society for Music Information Retrieval*. Porto, Portugal.

HathiTrust. (2015). Copyright. https://www.hathitrust.org/copyright (accessed 31 October 2015).

Schubert, P. and Cumming, J. (2015). "Another Lesson from Lassus: Quantifying Contrapuntal Repetition in the Duos of 1577." *Early Music* 43, no. 4 (September 2015).

Sigler, A., Wild, J. and Handelman, E. (2015). Schematizing the Treatment of Dissonance in 16th-Century Counterpoint. In *Proceedings of the Conference of the International Society for Music Information Retrieval*. Taipei, pp. 645–51.

Snydman, S., Sanderson, R., and Cramer, T. (2015). The International Image Interoperability Framework (IIIF): A community and technology approach for web-based images. In *Proceedings of the Archiving Conference*. Los Angeles, CA, 19–22 May.

Significance Testing for the Classification of Literary Subgenres

Lena Hettinger
lena.hettinger@uni-wuerzburg.de
University of Wuerzburg, Germany

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Wuerzburg, Germany

Isabella Reger
isabella.reger@uni-wuerzburg.de
University of Wuerzburg, Germany

Andreas Hotho
hotho@informatik.uni-wuerzburg.de
University of Wuerzburg, Germany

Introduction

The automatic classification of literary genres, especially of novels, has become a research topic in the last years (Underwood, 2014; Jockers, 2013). In the following we report on the results from a series of experiments using features like most frequent words, character tetragrams and different amounts of topics (LDA) for genre classification on a corpus of German novels. Two problems will be in the main focus of this paper and they are both caused by the same factor: The small number of labeled novels. So how can experiments be designed and evaluated reliably in a setting like this. We are especially interested in testing results for significance to get a better understanding of the reliability of our research. While statistical significance testing is quite established in many disciplines ranging from psychology to medicine, it is unfortunately not yet standard in digital literary studies.

The scarcity of labeled data is also one of the reasons some researchers segment novels. We will show that without a test for significance it would be easy to misunderstand our results and we will also show that using segments of the same novel in the test and the training data leads to an overestimation of the predictive capabilities of the approach.

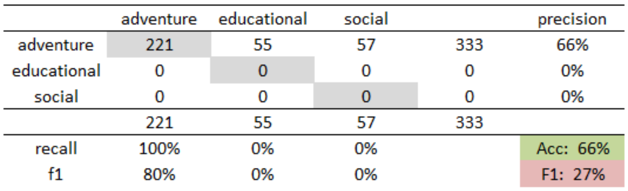
Setting

In the following we will describe our corpus and feature sets. Our corpus consists of 628 German novels mainly from the 19th century obtained from sources like TextGrid Digital Library[1] or Projekt Gutenberg[2]. The novels have been manually labeled according to their subgenre after research in literary lexica and handbooks. The corpus contains 221 adventure novels, 57 social novels and 55 educational novels; the rest belongs to a different or more than one subgenre.

Features are extracted and normalized to a range of [0,1] based on the whole corpus consisting of 628 novels. We have tested several feature sets beforehand and found stylometric and topic based to be the most promising (c.f. Hettinger et al., 2015). To represent stylometric features we employ 3000 most frequent words (mfw3000) and top 1000 character tetragrams (4gram). Topic based features are created using Latent Dirichlet Allocation (LDA) by Blei et al. (2003). In literary texts topics sometimes represent themes, but more often they represent topoi, often used ways of telling a story or parts of it (see also Underwood, 2012; Rhody, 2012). For each novel we derive a topic distribution, i.e. we calculate how strongly each topic is associated with each novel. We try different topic numbers and build ten models for each setting to reduce the influence of randomness in LDA models. We remove a set of predefined stop words as well as Named Entities from the novels as we have shown before that the removal of Named Entities tends to improve results.

Evaluation

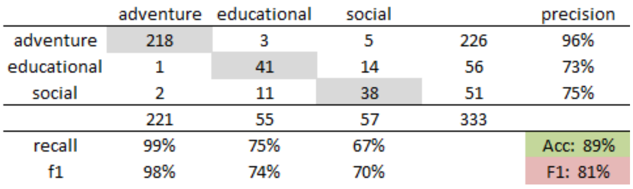
Classification is done by means of a linear Support Vector Machine (SVM) as we have already shown in Hettinger et al. (2015) that it works best in this setting (see also Yu, 2008). In each experiment we apply stratified 10-fold cross validation to the 333 labeled novels and report overall accuracy and F1-Score (c.f. Jockers, 2013). The majority vote (MV) baseline for our genre distribution yields an accuracy score of 0.66 and F1 score of 0.27 (see fig. 1).



	adventure	educational	social		precision
adventure	221	55	57	333	66%
educational	0	0	0	0	0%
social	0	0	0	0	0%
	221	55	57	333	
recall	100%	0%	0%		Acc: 66%
f1	80%	0%	0%		F1: 27%

Fig. 1: Cross table for majority vote baseline

In the cross tables of Figure 1 and 2 each column represents the true class and each row the predicted genre. Correct assignments are shaded in grey, average accuracy in green and average F1 score in red.



	adventure	educational	social		precision
adventure	218	3	5	226	96%
educational	1	41	14	56	73%
social	2	11	38	51	75%
	221	55	57	333	
recall	99%	75%	67%		Acc: 89%
f1	98%	74%	70%		F1: 81%

Fig. 2: Cross table for mfw 3000 as an example for classification results

Because there are not many labeled novels for genre classification we expanded our corpus by splitting every novel into ten equal segments. Features are then constructed independently for the resulting 3330 novel segments. To test the influence of the LDA topic parameter *t* in conjunction with having more LDA documents we evaluate topic features for *t* =100, 200, 300, 400, 500 (see figure 3 and 4).

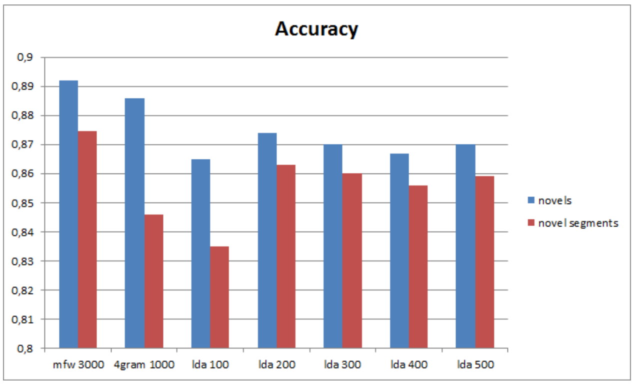


Fig. 3: Accuracy scores for novels and novel segments and different feature sets

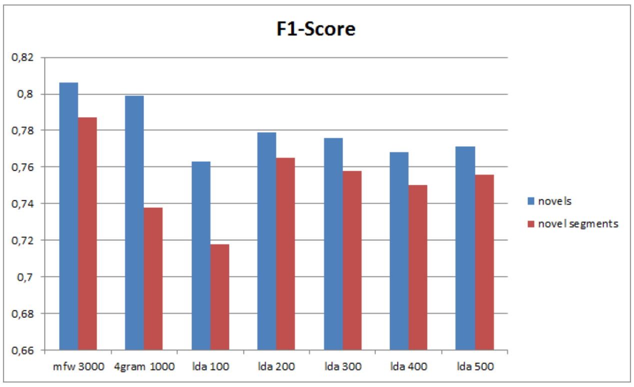


Fig. 4: F1 scores for novels and novel segments and different feature sets

Results show that our evaluation metrics tend to drop if novels are segmented. This could mean that genre is indeed a label for the whole literary work and not parts of it. On the other hand many differences are pretty small. Therefore we would like to test if these differences are statistically significant or if they should be attributed to chance.

Tests of statistical significance

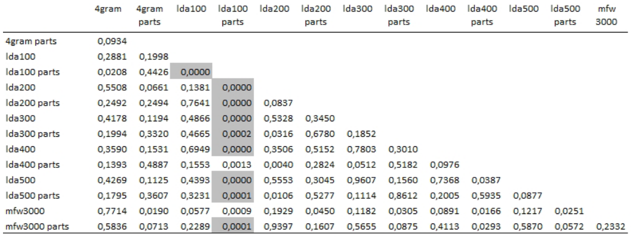
When working with literary corpora there are few genre labels available for two reasons. First, the task of labeling the genre of a novel is strenuous; second, literary studies have mostly concentrated on a rather small sample, the canonical novels. Another issue is the creation of a balanced corpus, because for historical reasons the distribution of literary genres is not uniform and also the process of selecting novels for digitization has made the situation even more complicated. This generally results in data sets of less than 1000 items or even less than 100, see for example Jockers (2013) where 106 novels form a corpus or Hettinger et al. (2015) where we evaluate on only 32 novels.

The problem arising from small corpora is that small differences in results may originate from chance. This can be investigated by using statistical tests (c.f. Kenny, 2013; Nazar and Sánchez Pol, 2006). A standard tool to detect if two data sets are significantly different is Student's t-test which we will use in the following to control the results of our experiments.

We use two variations of Student's t-test with $\alpha = 0.05$:

- the one-sample t-test to compare the accuracy of a feature set against the baseline
- the two-sample t-test to compare accuracy results for two feature sets

In both cases the data set considered consists of ten accuracy results from ten-fold cross validation and accordingly 100 data points for LDA from its ten models. Due to the small sample size we drop the assumption of equal variance for the two-sample t-test. The results for the one-sample t-tests show that every single feature set yields significantly better accuracy than the baseline (66.4%). We can therefore conclude that feature sets classify novels not randomly and that they do incorporate helpful genre clues.



	4gram	4gram parts	lda100	lda100 parts	lda200	lda200 parts	lda300	lda300 parts	lda400	lda400 parts	lda500	lda500 parts	mfw 3000
4gram parts	0,0934												
lda100	0,2881	0,1998											
lda100 parts	0,0208	0,4426	0,0000										
lda200	0,5508	0,0661	0,1381	0,0000									
lda200 parts	0,2492	0,2494	0,7641	0,0000	0,0837								
lda300	0,4178	0,1194	0,4866	0,0000	0,5328	0,3450							
lda300 parts	0,1994	0,3320	0,4665	0,0002	0,0316	0,6780	0,1852						
lda400	0,3590	0,1531	0,6949	0,0000	0,3506	0,5152	0,7803	0,3010					
lda400 parts	0,1393	0,4887	0,1553	0,0013	0,0040	0,2824	0,0512	0,5182	0,0976				
lda500	0,4269	0,1125	0,4393	0,0000	0,5553	0,3045	0,9607	0,1560	0,7368	0,0387			
lda500 parts	0,1795	0,3607	0,3231	0,0001	0,0106	0,5277	0,1114	0,8612	0,2005	0,5935	0,0877		
mfw3000	0,7714	0,0190	0,0577	0,0009	0,1929	0,0450	0,1182	0,0305	0,0891	0,0166	0,1217	0,0251	
mfw3000 parts	0,5836	0,0713	0,2289	0,0001	0,9397	0,1607	0,5655	0,0875	0,4113	0,0293	0,5870	0,0572	0,2332

Fig. 5: P-values for two sided t-test with $\alpha = 0.05$ on accuracy of genre classification using 333 German novels

P-values for the two-sided t-tests are reported in Figure 5. Due to the large number of tests we apply Holm-Bonferroni correction; the resulting statistically significant outcomes are shaded in grey. From Figure 5 it follows that differences between segmented and not-segmented novels are **not** statistically significant in most cases except for LDA with $t = 100$. Besides results do not differ significantly for different topic numbers $t = 100, 200, 300, 400, 500$.

An important assumption of the two-sample t-test is that both samples have to be independent. This is the case here as each time we do a cross validation we split the data independently from any other cross validation run. Thus, even if we repeat our experiments for a number of iterations (see e.g. Hettinger et al., 2015) we still get independent evaluation scenarios. Therefore we can apply the two-sided t-test in our setting to support our claims. In case of dependency of samples we could instead use paired t-tests on accuracy per novel.

Novel segmentation

A crucial factor when segmenting novels is how to distribute the segments between test and training data set. We decided that in our case we have to put all of the ten segments a novel was divided into either in the test or in the training data set as we want to derive the genre of a novel not seen before. Another possibility which Jockers (2013) exploited is to distribute segments randomly between training and test set. In his work "Macroanalysis" Jockers investigates how function words can be used to research aspects of literary history like author, genre etc. In the following we want to replicate the part concerning genre prediction using German novels.

When segments of one novel appear in both test and training data we achieve an accuracy of 97.5% and F1 score of 95.9% - that is close to perfect (see fig. 6). Such a partitioning of the novels dramatically overestimates predictive performance on unseen texts. In comparison, Jockers (2013) achieves an average F1 score of 67% on twelve genre classes. His results are worse because we are only using three different genres while he is doing a multiclass classification with 12 classes. But nevertheless 67% probably still overestimates the real predictive power of this approach, because in our setup using the segments in both, test and training data, increased F1 by more than 17%.

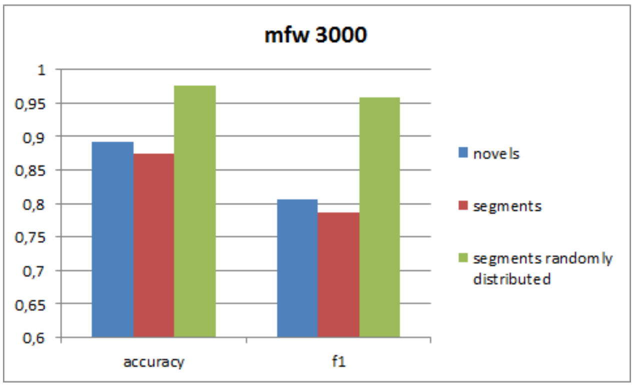


Fig. 6: Results for different partitioning strategies

Conclusion

In this work we looked at the methodology and evaluation of genre classification of German novels and discussed some of the methodical pitfalls of working with data like this. We discovered that only some of our results turned out to be statistically significant whereas for example the statement, that stylometric perform better than topic-based features, could not be fortified. Therefore our opinion is that research findings on small data sets should be scrutinized especially carefully for example by using statistical tests.

Bibliography

Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, **3**: 993-1022.

Hettinger, L. et al. (2015). Genre Classification on German Novels, *Proceedings of the 12th International Workshop on Text-based Information Retrieval,* Valencia, Spain.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Illinois: University of Illinois Press.

Kenny, A. (1982). *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. New York: Elsevier.

Nazar, R. and Sánchez Pol, M. (2006). An Extremely Simple Authorship Attribution System, *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics/Language and the Law*, Barcelona, Spain, 2006.

Rhody, L. M. (2012). Topic Modeling and Figurative Language, *Journal of Digital Humanities*, **2**(1). http://journalofdigitalhumanities.org/2-1/ (accessed 1 November 2015).

Underwood, T. (2012). Topic Modeling Made Just Simple Enough, *Blog post 7 April 2012*. http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ (accessed 1 November 2015).

Underwood, T. (2014). Understanding Genre in a Collection of a Million Volumes, *Interim Report*. http://dx.doi.org/10.6084/m9.figshare.1281251 (accessed 26 August 2015).

Yu, B. (2008). An Evaluation of Text Classification Methods for Literary Study, *Literary and Linguistic Computing* **23**: 327-43.

Notes

[1] textgrid.de/digitale-bibliothek

[2] gutenberg.spiegel.de

Microanalyzing Parts of Texts

David L. Hoover
david.hoover@nyu.edu
New York University, United States of America

Some broad questions require the analysis of huge collections of texts. Other broad questions and many narrower ones require microanalyzing parts of texts. Some microanalyses are unproblematic: narrative structure and its relationship to chapter divisions can be studied simply by dividing texts into chapters. Analyzing narrative or dialogue only, or the relationships between these and chapter divisions, may be much more problematic, as may analyzing a novel that also contains letters, diaries, legends, and poetry. Some or all of these may be more appropriately analyzed separately or ignored. Difficulties multiply for multiple narrators whose narratives contain dialogue and subdivisions.

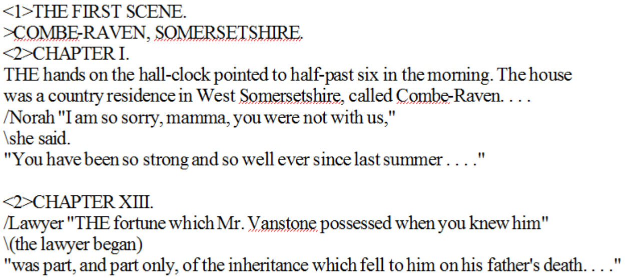
One of the most difficult tasks is analyzing the character dialogue in a novel. Burrows showed that the frequencies of the most frequent words in their dialogue can distinguish Jane Austen's characters from each other (1987), but few scholars have followed his lead, at least partly because of the tedium and difficulty of separating character parts. McKenna and Antonia (1996) were an early exception, but most related work involves epistolary novels or multiple narrators, where the separation of parts is simpler (Stewart 2003; Rybicki 2006; Ramsay 2011; Balossi 2014; Hoover, Culpeper, and O'Halloran 2014; Hoover 2010, and forthcoming).) Consider the case of Sherlock Holmes. Perhaps, as Moretti argues, "Doyle owes his phenomenal success to his greater skill in the handling of clues" (2004, 48), but Holmes and Watson are also extraordinarily fascinating characters. Analyzing their voices for distinctiveness requires comparing them with his other characters. Because reliable results require substantial amounts of text, I focus here on the longest Holmes novel, *The Hound of the Baskervilles* (*Hound*, below).

Extracting the dialogue computationally still requires the tedious and error-prone manual separation of the character parts and identification of the speakers. Typically characters are too numerous to open separate dialogue files for all of them, and multiple files increase copying and pasting errors. Initial decisions about the handling of dialogue may also change, requiring painstaking re-editing. Instead, I introduce very simple markup that is then processed in "Analyze Textual Divisions," an Excel spreadsheet with macros. The markup, powerful enough for texts with quite complex structures, is also simple, flexible, and customizable:

<1> text division 1
<2> text division 2
<3> text division 3
<4> text division 4
[ ] Letter writer
{ } Letter addressee
/ speaker
\ speech marker
> copy without processing
^ special character follows

For Wilkie Collins's complex novel *No Name,* with scenes containing chapters, which contain letters and other documents, the four divisions are "Scene", "Chapter", "Letter", and "Document." (The spreadsheet includes brief excerpts from this novel with mark-up.) Epistolary novels might use "Letter," and others might use "Volume" and "Book." For texts with multiple narrators and for plays "Narrator" and "Act" and "Scene" are obvious divisions. The top-level division, like the rest of the markup, can be modified. For *No Name*, division one is defined as follows: div1name = "Scene". Novels divided into books could use "div1name = "Book." Alternatively, after the macro operates, the labels can be changed as desired.

Here is a truncated version of *No Name*:

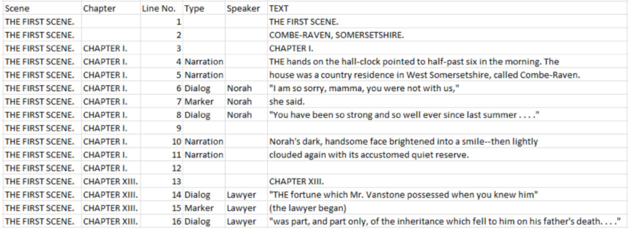


```
<1>THE FIRST SCENE.
>COMBE-RAVEN, SOMERSETSHIRE.
<2>CHAPTER I.
THE hands on the hall-clock pointed to half-past six in the morning. The house
was a country residence in West Somersetshire, called Combe-Raven. . . .
/Norah "I am so sorry, mamma, you were not with us,"
\she said.
"You have been so strong and so well ever since last summer . . . ."

<2>CHAPTER XIII.
/Lawyer "THE fortune which Mr. Vanstone possessed when you knew him"
\(the lawyer began)
"was part, and part only, of the inheritance which fell to him on his father's death. . . ."
```

A "<1>" has been inserted to mark "THE FIRST SCENE." as division one, and all lines in the first scene will be so labeled. In line two, ">" indicates that "COMBE-RAVEN, SOMERSETSHIRE.," which seems like a scene-setting label, not narrative, should not be processed (epigraphs or poems might be treated similarly). In line three, "CHAPTER I." marks division two. In line six, "/Norah" labels lines 6-8 as hers (the person addressed could, like a letter addressee, be marked with {}). In line seven, "\she said" is a speech marker, categorized separately because they sometimes vary interestingly and because "she said" seems to me neither dialogue nor narration. In line eight, the quotation mark indicates dialogue. The blank ninth line changes the label from dialogue to narrative until marked otherwise. The beginning of chapter thirteen is marked similarly. Later in the novel, embedded letters are marked with "Letter writer" and "Letter addressee." Finally, "^" must begin any line that would otherwise begin with "+", "-", or "=" (reserved characters in Excel). (Line-division can be changed instead, except where required line breaks force special characters to the beginning.)

With the Analyze Textual Divisions spreadsheet and the marked-up text open in Excel, the macro processes the text line by line, producing the results below (the marked-up text and empty columns have been deleted). Each line gets a scene label, and, beginning in line three, a chapter label, and all the lines are numbered. Lines 4-5 are marked as narration, lines 6 and 8 as Dialogue, and line 7 as Marker, and the speaker is entered for lines 6-8 and 14-16. The processed text appears on the right with all markup removed.

The text could be marked up in TEI and the character parts extracted with XSLT, but the markup here is much simpler and easier to learn, and the spreadsheet has advantages over XSLT. Excel's built-in sorting function can handle several levels of sorting, for example, so that the dialogue can be sorted by type, scene, chapter, speaker, and line number, all at once. The unmarked processed text, after sorting, can be divided and analyzed however the analyst desires with plain-text tools. Sorting on the line number restores the original order for further analysis, and errors can be corrected in the original text, and the analysis re-run. (See my Excel Text-Analysis Pages at http://wp.nyu.edu/exceltextanalysis/ for detailed instructions.) This method works especially well for short, simple texts like *Hound,* with character parts too short to be analyzed by chapter; the dialogue can be marked with just speaker and speech marker characters, and > and ^.



Scene	Chapter	Line No.	Type	Speaker	TEXT
THE FIRST SCENE.		1			THE FIRST SCENE.
THE FIRST SCENE.		2			COMBE-RAVEN, SOMERSETSHIRE.
THE FIRST SCENE.	CHAPTER I.	3			CHAPTER I.
THE FIRST SCENE.	CHAPTER I.	4	Narration		THE hands on the hall-clock pointed to half-past six in the morning. The
THE FIRST SCENE.	CHAPTER I.	5	Narration		house was a country residence in West Somersetshire, called Combe-Raven.
THE FIRST SCENE.	CHAPTER I.	6	Dialog	Norah	"I am so sorry, mamma, you were not with us,"
THE FIRST SCENE.	CHAPTER I.	7	Marker	Norah	she said.
THE FIRST SCENE.	CHAPTER I.	8	Dialog	Norah	"You have been so strong and so well ever since last summer . . . ."
THE FIRST SCENE.	CHAPTER I.	9			
THE FIRST SCENE.	CHAPTER I.	10	Narration		Norah's dark, handsome face brightened into a smile--then lightly
THE FIRST SCENE.	CHAPTER I.	11	Narration		clouded again with its accustomed quiet reserve.
THE FIRST SCENE.	CHAPTER I.	12			
THE FIRST SCENE.	CHAPTER XIII.	13			CHAPTER XIII.
THE FIRST SCENE.	CHAPTER XIII.	14	Dialog	Lawyer	"THE fortune which Mr. Vanstone possessed when you knew him"
THE FIRST SCENE.	CHAPTER XIII.	15	Marker	Lawyer	(the lawyer began)
THE FIRST SCENE.	CHAPTER XIII.	16	Dialog	Lawyer	"was part, and part only, of the inheritance which fell to him on his father's death. . . ."

To test the distinctiveness of the character voices in *Hound*, I selected all character parts at least 1,500 words long, and divided longer parts into 1,500-word sections. Initial testing was disappointing. Although the sections of dialogue by Stapleton, Mortimer, and Watson grouped correctly, those by Baskerville, Barrymore, and Holmes did not, casting doubt on the distinctiveness of their voices. The section of Baskerville's dialogue that groups with Barrymore's, however, consists almost entirely of a conversation between the two, so that similarity of topic may skew the results. More significantly, the first six sections of Holmes's dialogue consistently group correctly. The final two, which consist almost entirely of the final chapter, and which tend to group separately from all others, are Holmes's explanation of the case to Watson. Nominally dialogue, this chapter is more like narration, a genre difference that is almost certainly responsible for the anomalous clustering. Removing the final chapter and sorting the lines of Baskerville's dialogue in random order to blunt any topical or thematic effects produces the cluster analysis shown in Fig. 1, based on the 225mfw (most frequent words).

Cluster analysis is an exploratory statistical method that compares the frequencies of a set of words across a set of texts to determine which texts use those words at the most similar frequencies. The nearer to the left that sections join together into a single cluster, the more similarly they use the words. All sections in Fig. 1 group correctly by speaker, and several sections of Holmes's dialogue are the most similar, and the results are correct across analyses based on the 125-325mfw. Doyle's use of clues may have helped the Sherlock Holmes stories succeed, but the distinct character voices also seem likely to be a factor. (The analysis here uses Ward Linkage and squared Euclidean distance; the often-used complete linkage gives weaker results.)

Separating character dialogue can never be easy, but my spreadsheet makes it much easier. It also provides a versatility in comparing multiple kinds of textual divisions that may encourage more in-depth analysis of dialogue and characterization and enhance our understanding of how texts work.

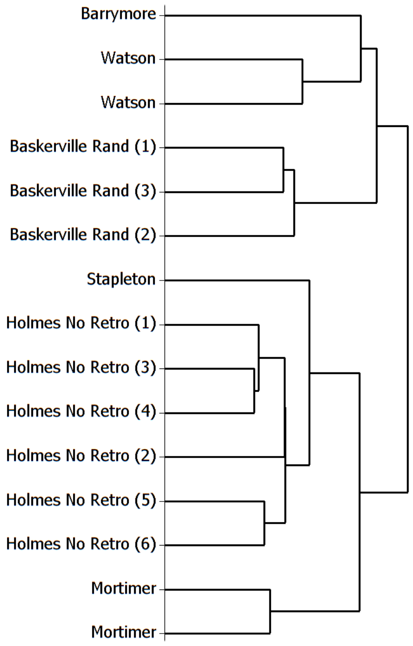


Fig. 1: Character Dialogue in *Hound* (225mfw)

Bibliography

Balossi, G. (2014). *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves.* Amsterdam: Johns Benjamins.

Burrows, J. (1987). *Computation into Criticism.* Oxford: Clarendon Press.

Hoover D. (forthcoming). Argument, evidence, and the limits of digital literary studies. In Gold, M. (ed), *Debates in the Digital Humanities,* University of Minnesota Press.

Hoover D. (2010). Some approaches to corpus stylistics. In Yu Dongmin (ed), *Stylistics: Past, Present and Future*. Shanghai Foreign Language Education Press, pp. 40-63.

Hoover, D., Culpeper, J., and O'Halloran, K. (2014). *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. London: Routledge.

McKenna, C. W. F., and Antonia, A. (1996). 'A few simple words' of interior monologue in "Ulysses": reconfiguring the evidence. *Literary and Linguistic Computing*, **11**(2): 55-66.

Moretti, F. (2004). Graphs, maps, trees: abstract models for literary history — 3, *New Left Review*, **28**: 43-63.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

Rybicki, J. (2006). Burrowing into translation: character idiolects in Henryk Sienkiewicz's trilogy and its two English translations. *Literary and Linguistic Computing*, **21**(1): 91-103.

Stewart, L. (2003). Charles Brockden Brown: quantitative analysis and literary interpretation. *Literary and Linguistic Computing*, **18**(2): 129-38.

Textual Variation, Text-Randomization, and Microanalysis

David L. Hoover
david.hoover@nyu.edu
New York University, United States of America

Computational stylistics often analyzes style variation, including chronological change and the dialogue or narration of multiple characters or narrators (Craig 1999; Stamou 2008; McKenna and Antonia 1996; Stewart 2003; Burrows 1987; Hoover 2003, 2007). Here I suggest it is sometimes desirable to omit parts of texts, or to randomize a text to mask some variation, either analyzing parts of texts in random order or truncating such parts to create a word list based on equal amounts of text (see also Burrows, 1992).

Analyzing Doyle's *The Hound of the Baskervilles,* shows that two sections of Holmes's dialogue and one of Baskerville's fail to cluster correctly, but both failures likely result from intra-textual variation that should *not* influence the analysis. Holmes's outlier sections are a retrospective explanation more like narration than dialogue. This genre difference disrupts Holmes's otherwise strongly consistent voice. I have similarly suggested ignoring the final "summing up" chapter of *The Waves* when analyzing its six narrative voices (Hoover forthcoming). Baskerville's problematic section is largely a conversation with Barrymore, which may cause their voices to merge and their sections to cluster.

Special pleading in the face of apparent failure seems potentially illegitimate, but removing Holmes's retrospective "dialogue" leaves the rest consistently clustered. Furthermore, Holmes's dialogue from other stories clusters perfectly with his normal dialogue from *Hound*, so that inter-textual consistency supports an intra-textual argument. Retesting *Hound* with Baskerville's dialogue randomly sorted clusters his three sections separately from Barrymore's section, which forms its own cluster. But such randomizing to dampen intra-textual variation seems questionable. Will less distinctive voices cluster with these homogenized parts? Will randomization produce specious clustering of sections that are not "really" consistent in style?

Consider the nine fairly distinct main narrators in Collins's *The Moonstone* (Hoover, Culpeper, and O'Halloran, 2014). Here, randomizing the parts greatly improves accuracy, correctly clustering all sections by all narrators over a wide range of analyses. Normal analysis of Faulkner's *As I Lay Dying* clusters Darl's and Vardaman's sections correctly and clusters all of Tull's sections with Cash's section over a wide range of analyses (Hoover, 2010), but randomizing each character's narrative correctly clusters all the characters, even when divided into much shorter sections.

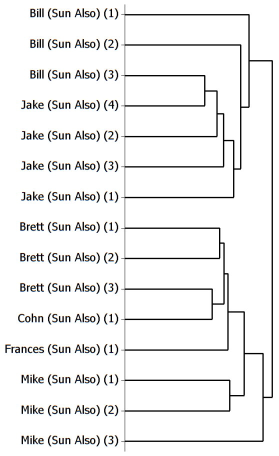


Fig. 1. Character Dialogue in *The Sun Also Rises*–Standard Analysis, 700MFW

What about less distinct character parts? Randomization only slightly improves results for the letter writers in *The Coquette* (an epistolary novel), showing that randomization does not always produce artificial consistency. Jake and Brett from Hemingway's *The Sun Also Rises* are memorable characters, yet their dialogue does not cluster very consistently in standard analyses, never approaching the success with Collins's or Faulkner's narrators . Figure 1 shows the most accurate clustering, based on the 700 MFW (most frequent words); all other analyses are weaker. Randomizing the character parts produces the completely correct results shown in Fig. 2 for analyses based on the 400-800MFW. (Cluster analysis is an exploratory statistical method that compares the frequencies of a set of words across a set of texts to determine which texts use those words at the most similar frequencies. The further to the left that two or more texts form a cluster, the more similar they are.) Randomization transforms the poor character separation in a standard analysis of James's *The Ambassadors* into a clear separation of all except Bilham, in analyses based on the 500-800MFW. The same is true for *Jane Eyre*, though Jane's dialogue seems problematic, as her story ranges from childhood to adulthood in five different settings. Many more texts will have to be analyzed to determine why randomization produces varying results and where it is appropriate.

Dracula reacts similarly. A standard analysis, shown in Fig. 3, fails to cluster all the sections correctly. In the middle of the large lower cluster is a problematic mixed cluster containing Van Helsing's single section and the final sections from Mina, Harker, and Seward. These sections narrate the race to capture Dracula at the end of the novel, and this provides an opportunity to push the question of randomization further. Because the *Dracula* narratives range widely in length, the full word list is skewed away from Lucy (4,400) and Van Helsing (5,200) and toward Mina (22,000), Harker 1 (19,000), Harker 2 (14,000) and Seward (35,000). A word list based on the first 6,000 words from the randomized narratives of these five characters (half from each of Harker's) clusters the narrative much more accurately and clearly distinguishes the narratives of Lucy and Van Helsing (see Fig. 4).

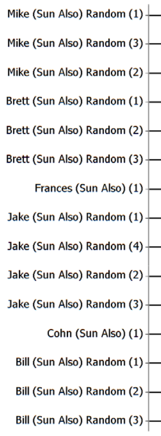

Mike (Sun Also) Random (1)
Mike (Sun Also) Random (3)
Mike (Sun Also) Random (2)
Brett (Sun Also) Random (1)
Brett (Sun Also) Random (2)
Brett (Sun Also) Random (3)
Frances (Sun Also) (1)
Jake (Sun Also) Random (1)
Jake (Sun Also) Random (4)
Jake (Sun Also) Random (2)
Jake (Sun Also) Random (3)
Cohn (Sun Also) (1)
Bill (Sun Also) Random (1)
Bill (Sun Also) Random (2)
Bill (Sun Also) Random (3)

Fig. 2. Character Dialogue in *The Sun Also Rises*–Randomized Parts, 700MFW

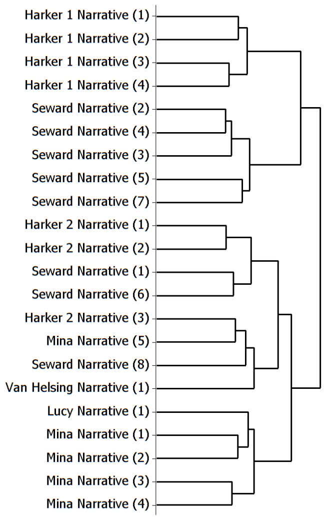


Fig. 3. Six Dracula Narrators–Standard Analysis, 900MFW

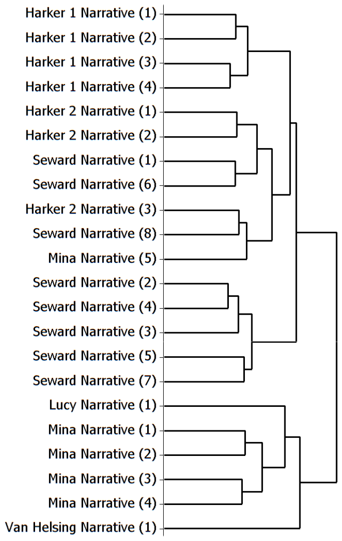


Fig. 4. Six Dracula Narrators–Word List Based on 6,000 Randomized Words, 900MFW

Harker's, Mina's, and Seward's final sections remain clustered in Fig. 4, however. Analyzing these and Van Helsing's in shorter sections, using a word list based on the equalized randomized parts, again improves results. Seward's and Van Helsing's sections cluster separately, as do Mina's second, third, and fourth; only her first section clusters with Jonathan's. Thus the voices remain relatively distinct, though inflected by the effects of the narrative structure. Because Mina's wayward section is mainly a memo, its failure to cluster with the sections from her journal may also be explicable. Further analysis of the sub-genres of *Dracula* should provide additional insights into intra-textual variation.

Consider now a very different kind of text posing different problems–a collection of nearly 400 high-stakes writing exams administered in U.S. high schools. These very short texts (128-1307 words) in multiple genres were written in response to a wide variety of prompts from multiple states (Jeffrey 2010; Jeffrey, Hoover, and Han 2013). This makes testing for the characteristic vocabulary of high- and low-scoring texts very challenging. Here I use a variant of Zeta, which identifies words used consistently by one group and avoided consistently by another (Burrows, 2006; Craig and Kinney, 2009; Hoover, 2013). Because the entire vocabulary is being tested, a normal word list seems appropriate. I also combine the low-scoring texts into one large text and the high-scoring texts into another, randomize the lines of each, and then compare analyses based on sections of the combined texts with analyses based on the combined and randomized texts. Holding out twenty high- and twenty low-scoring sections for testing, I use the eighty-six remaining sections for training. The combined texts distinguish high and low test texts fairly well (Fig. 5), but the combined-randomized texts greatly improve the results (Fig. 6). (The vertical and horizontal axes record the percentage of word types in each section that are characteristic of low-scoring and high-scoring texts, respectively.)

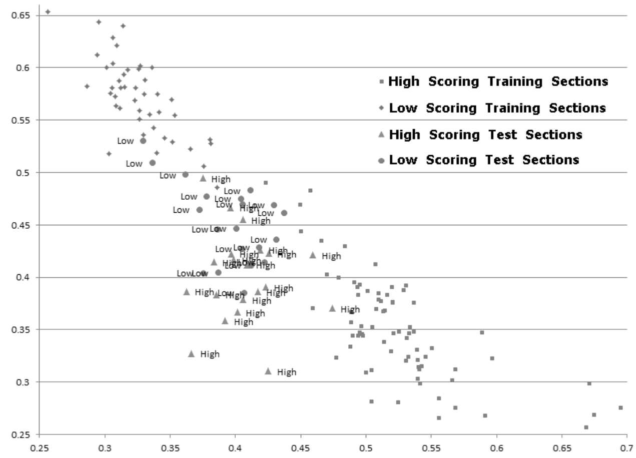


Fig. 5. Combined High- and Low-Scoring Exit Texts

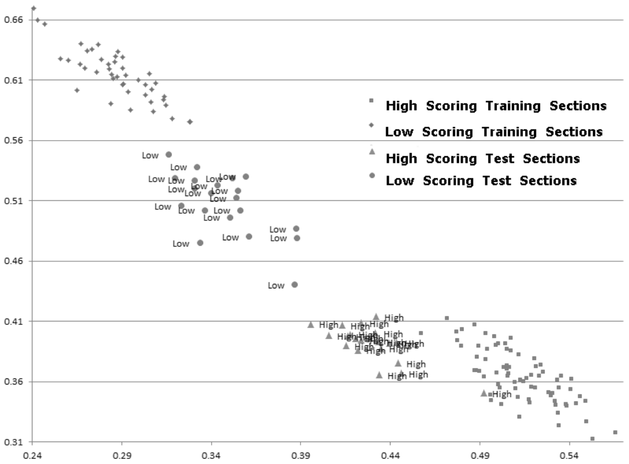


Fig. 6. Combined-Randomized High- and Low-Scoring Exit Texts

Certainly (intra-)textual variation helps to create memorable characters and narrators, and it is often crucially linked to description or narrative action. Some kinds of variation, however, can mask important kinds of consistency and unity, and a reasoned argument can be made for ignoring some sections of texts, or for analyzing them in randomized form. Alternatively, or in addition, word lists based on truncated and randomized sections also often improve results. Though the value and legitimacy of such transformations and truncations will need further study, they seem a promising line of research.

Bibliography

Burrows, J. (2006). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, **22**(1): 27-47.

Burrows, J. (1992). Not unless you ask nicely: the interpretative nexus between analysis and information, *Literary and Linguistic Computing*, **7**(2): 91-109.

Burrows, J. (1987). *Computation into Criticism.* Oxford: Clarendon Press.

Craig, H. (1999). Contrast and change in the idiolects of Ben Jonson characters. *Computers and the Humanities*, **33**(3): 221-40.

Craig, H, and Kinney, A. (Eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge U Press.

Hoover D. (forthcoming). Argument, evidence, and the limits of digital literary studies. In Gold, M. (Ed), *Debates in the Digital Humanities.* University of Minnesota Press.

Hoover, D. (2013). The full-spectrum text-analysis spreadsheet, *Digital Humanities 2013: Conference Abstracts.* Lincoln, NE: Center for Digital Research in the Humanities, University of Nebraska, pp. 226-29.

Hoover, D. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style*, **41**(2): 160-89.

Hoover, D. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, **18**(4): 341-60.

Hoover, D., Culpeper, J. and O'Halloran, K. (2014). *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. London: Routledge.

Jeffrey, J. (2010). Voice, genre, and intentionality: an integrated methods study of voice criteria examined in the context of large scale-writing assessment. Diss. English Education. New York University.

Jeffrey, J., Hoover, D. and Han, M. (2013). Lexical variation in highly and poorly rated US secondary students' writing: implications for the common core writing standards, *AERA 2013 Annual Meeting*, San Francisco, April 27-May 1.

McKenna, C. W. F. and Antonia, A. (1996). A few simple words' of interior monologue in "Ulysses": reconfiguring the evidence. *Literary and Linguistic Computing*, **11**(2): 55-66.

Stamou, C. (2008). Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, **23**(2): 181-99.

Stewart, L. (2003). Charles Brockden Brown: quantitative analysis and literary interpretation. *Literary and Linguistic Computing*, **18**(2): 129-38.

Assessing a Shape Descriptor for Analysis of Mesoamerican Hieroglyphics: A View Towards Practice in Digital Humanities

Rui Hu
rhu@idiap.ch
Idiap Research Institute, Switzerland

Jean-Marc Odobez
odobez@idiap.ch
Idiap Research Institute, Switzerland; Ecole Polytechnique Federale de Lausanne (EPFL)

Daniel Gatica-Perez
gatica@idiap.ch
Idiap Research Institute, Switzerland; Ecole Polytechnique Federale de Lausanne (EPFL)

Introduction

Technological advances in digitization, automatic image analysis and information management are enabling the possibility to analyze, organize and visualize large cultural datasets. As one of the key visual cues, shape feature has been used in various image analysis tasks such as handwritten character recognition (Fischer, 2012; Franken, 2013), sketch analysis (Eitz, 2012), etc. We assess a shape descriptor, within the application domain of Maya hieroglyphic analysis. Our aim is to introduce this descriptor to the wider Digital Humanities (DH) community, as a shape analysis tool for DH-related applications.

The Maya civilization is one of the major cultural developments in ancient Mesoamerica. The ancient Maya language infused art with uniquely pictorial forms of hieroglyphic writing, which represents an exceptionally rich legacy (Stone, 2011). Most Maya texts were written during the Classic period (AD 250-900) of the Maya civilization on various media types, including stone monuments. A special class of Maya texts was written on bark cloths as folding books from the Post-Classic period (AD 1000-1519). Only three such books (namely the Dresden, Madrid and Paris codices) are known to have survived the Spanish Conquest. A typical Maya codex page contains icons, main sign glyph blocks, captions, calendric signs, etc. Fig. 1 illustrates an example page segmented into main elements (Gatica-Perez, 2014). In this paper, we are interested in the main signs.

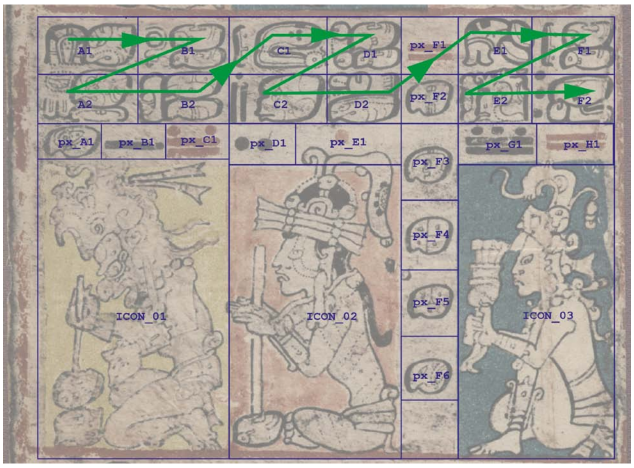


Figure 1. Page 6b of the Dresden Codex, showing individual constituting elements framed by blue rectangles (Hu, 2015), Green arrows indicate reading order of the main sign blocks, generated by Carlos Pallan based on SLUB (http://digital.slub-dresden.de/werkansicht/dlf/2967/1/) online open source image.

Maya hieroglyphic analysis requires epigraphers to spend a significant amount of time browsing existing catalogs to identify individual glyphs. Automatic Maya glyph analysis has been addressed as a shape matching problem, and a robust shape descriptor called Histogram of Orientation Shape Context (HOOSC) was developed in (Roman-Rangel, 2011). Since then, HOOSC has been successfully applied for automatic analysis of other cultural heritage data, such as Oracle-Bones Inscriptions of ancient Chinese characters (Roman-Rangel, 2012), and ancient Egyptian hieroglyphs (Franken, 2013). It has also been applied for generic sketch and shape image retrieval (Roman-Rangel, 2012). Our recent work extracted statistic Maya language model and incorporated it for glyph retrieval (Hu, 2015).

The goal of this paper is two-fold:

1. Introduce the HOOSC descriptor to be used in DH-related shape analysis tasks (code available at: http://www.idiap.ch/paper/maaya/hoosc/);

2. Discuss key issues for practitioners, namely the effect that certain parameters have on the performance of the descriptor. We describe the impact of such choices on different data types, specially for 'noisy' data as it is often the case with DH image sources.

Automatic Maya Hieroglyph Recognition

We conduct glyph recognition with a retrieval system proposed in (Hu, 2015). Unknown glyphs are considered as queries to match with a database of known glyphs (retrieval database). Shape and context information are considered. Fig.2 illustrates a schema of our approach. We study the effect of different HOOSC parameter choices on the retrieval results.

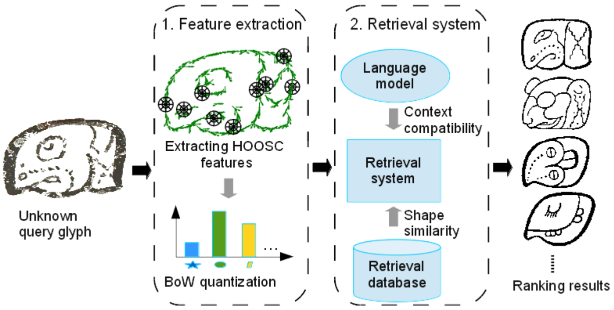


Figure 2. Retrieval system pipeline.

Datasets

We use three datasets, namely the 'Codex', 'Monument' and 'Thompson'. The first two are used as queries to search within the retrieval database ('Thompson').

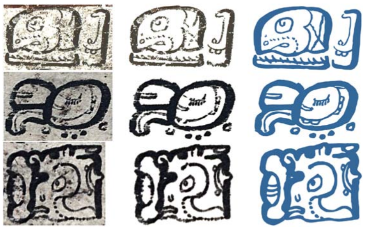


Figure 3. Digitization quality: (left) raw glyph blocks cropped from Dresden codex; (middle) clean raster images produced by removing the background noise; (right) reconstructed high-quality vectorial images.

The 'Codex' dataset contains glyph blocks from the three surviving Maya codices. See Fig.3 for examples. Glyph blocks are typically composed of combinations of individual signs. Fig.4 shows individual glyphs segmented from blocks in Fig.3. Note the different degradation levels across samples. We use two sub-datasets: 'codex-small', composed of 156 glyphs segmented from 66 blocks, for which we have both clean raster and high-quality reconstructed vectorial representations (see Fig.4) to study the impact of the different data qualities on the descriptor; and a 'codex-large' dataset, which is more extensive, comprising only the raster representation of 600 glyphs from 229 blocks.

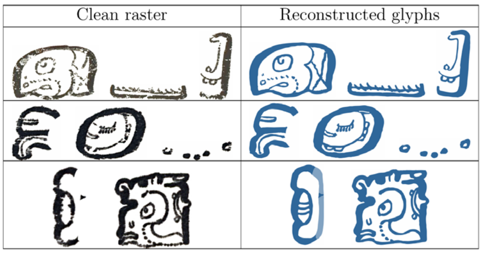


Figure 4. Example glyph strings generated from blocks shown in Figure 3.

The 'Monument' dataset is an adapted version of the Syllabic Maya dataset used in (Roman-Rangel, 2011), which contains 127 glyphs of 40 blocks extracted from stone monuments. It is a quite different data source to the codex data, in terms of historical period, media type, and data generation process. Samples are shown in Fig.5.

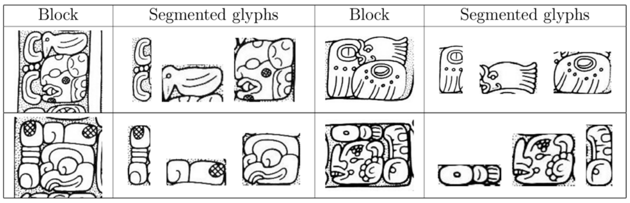


Figure 5. Example blocks and segmented glyph strings form the 'Monument' dataset.

To form the retrieval database ('Thompson'), we scanned and segmented all the glyphs from the Thompson catalog (Thompson, 1962). The database contains 1487 glyph examples of 892 different sign categories. Each category is usually represented by a single example image. Sometimes multiple examples are included; each illustrates a different visual instance or a rotation variant. Fig.6 shows glyph examples.

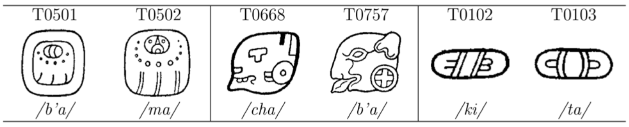


Figure 6. Thompson numbers, visual examples, and the syllabic values of glyph pairs. Each pair contains two different signs with similar visual features (Hu, 2015). All examples are taken from (Thompson, 1962).

Shape-based retrieval

Feature extraction and similarity matching are the two main steps for our shape-based glyph retrieval framework.

Glyphs are first pre-processed into thin lines. To do so, an input glyph (Fig.7(a)) is first converted into a binary shape (Fig.7 (b)). Thin lines (Fig.7(c)) are then extracted through mathematical morphology operations. Fig.7(c)-(d) show the high-quality reconstructed binary image, and the extracted thin lines.

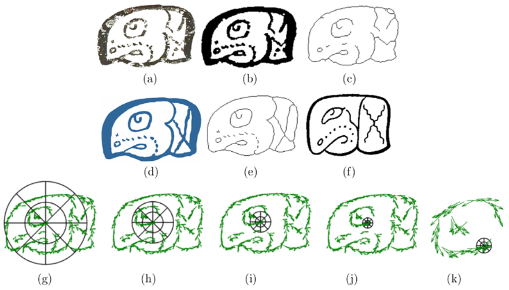


Figure 7. Extracting HOOSC descriptor: (a) input clean raster image; (b) binary image; (c) thinned edge of (b); (d) reconstructed vector representation of (a); (e) thinned edge of (d); (f) corresponding groundtruth image in the catalog; (g)-(k) spatial partition of a same pivot point with five different ring sizes (1, ½, ¼, 1/8, 1/16, all defined as a proportion to the mean of the pairwise distance between pivot points) on the local orientation field of the thinned edge image (c). Note that we zoomed in to show the spatial context of 1/16 in (k).

HOOSC descriptors are then computed at a subset of uniformly sampled pivot points along the thin lines. HOOSC combines the strength of Histogram of Orientation Gradient (Dalal, 2005) with circular split binning from the shape context descriptor (Belongie, 2002). Given a pivot point, the HOOSC is computed on a local circular space centred at the pivot's location, partitioned into rings and evenly distributed angles. Fig.7 (g)-(k) show different sizes of the circular space (referred to as spatial context) partitioned into 2 rings and 8 orientations. A Histogram-of-orientation-gradient is calculated within each region. The HOOSC descriptor for a given pivot is the concatenation of histograms of all partitioned regions.

We then follow the Bag-of-Words (BoW) approach, where descriptors are quantized as visual words based on the vocabulary obtained through K-means clustering on the set of descriptors extracted from the retrieval database. A histogram representing the count of each visual word is then computed as a global descriptor for each glyph. In all experiments, we use vocabulary size k=5000.

Each query is matched with glyphs in the retrieval database, by computing shape feature similarity using the L1 norm distance.

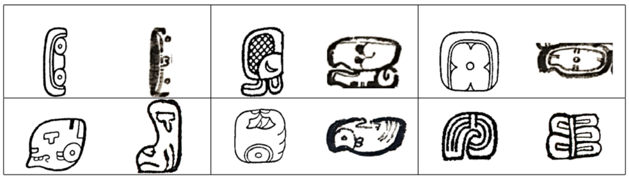


Figure 8. Six pairs of glyph signs (Hu, 2015). Each pair contains a query glyph from the 'Codex' dataset (right), and their corresponding groundtruth in the catalog (left).

Incorporating context information

Shape alone is often ambiguous to represent and distinguish between images. In the case of our data, different signs often share similar visual features (see Fig.6); glyphs of the same sign category vary with time, location, and the different artists who produced them (see Fig.8); additionally, surviving historical scripts often lose visual quality over time. Context information can be used to complement the visual features.

Glyph co-occurrence within single blocks encodes valuable context information. To utilize this information, we arrange glyphs within a single block into a linear string according to the reading order (see Fig.4 and Fig.5), and consider the co-occurrence of neighbouring glyphs using an analogy to a statistical language model. For each unknown glyph in the string, we compute its probability to be labelled as a given category by considering not only the shape similarity, but also the compatibility to the rest of the string.

We apply the two language models extracted in (Hu, 2015), namely the ones derived from the Maya Codices Database (Vail, 2013) and the Thompson catalog (Thompson, 1962), which we refer to as the 'Vail' and the 'Thompson' models. We use Vail model with smoothing factor α=0 for the 'Codex' data, and the Thompson model with α=0.2 for the 'Monument' data.

Experiments and Results

Our aim is to demonstrate the effect of various HOOSC parameters on retrieval results.

Experimental setting

We illustrate the effect of 3 key parameters:

- Size of the spatial context region within which HOOSC is computed

A larger region encodes more context information and therefore captures more global structure of the shape. However, in the case of image degradation, a larger region could contain more noise. We evaluate five different spatial contexts as shown in Fig.7(g)-(k). The circular space is distributed over 8 angular intervals.

- Number of rings to partition the local circular region

This parameter represents different partition details. We evaluate either 1 or 2 rings, the inner ring covers half the distance to the outer ring. Each region is further characterized by a 8-bin histogram of the local orientations.

- Position information

Relative position (i, j) of a pivot in the 2-D image plane can be concatenated to the corresponding HOOSC feature.

Results and discussion

Fig.9 shows the average groundtruth ranking in the retrieval results with different parameter settings, on three query sets, *e.g.* 'Codex-large', 'Codex-small' and 'Monument'. Each query image usually has only one correct match (groundtruth) in the retrieval database. The smaller the average ranking value, the better the result. From Fig.9 we can see the following:

- In most cases, the best results are achieved by using the largest spatial context, with finer partitioning details (2 rings in our case);
- When the location information is not considered, results show a general trend of improving with increasing ring sizes. However, the results are more stable when the position information is encoded, *e.g.* a smaller ring size can also achieve promising results when the location information is incorporated. This is particularly useful when dealing with noisy data, where a smaller ring size is preferred to avoid extra noise been introduced by a larger spatial context;
- The results do not benefit from a finer partition, when a small spatial context is considered. However, results improve with finer partitions, when the spatial context becomes larger.
- Position information is more helpful when a small spatial context is considered.

Fig.10 shows example query glyphs and their top returned results.

Conclusion

We have introduced the HOOSC descriptor to be used in DH-related shape analysis tasks. We discuss the effect of parameters on the performance of the descriptor. Experimental results on ancient Maya hieroglyph data from two different sources (codex and monument) suggest that a larger spatial context with finer partitioning detail usually leads to better results, while a smaller spatial context with location information is a good choice for noisy/damaged data. The code for HOOSC is available so DH researchers can test the descriptor for their own tasks.

Acknowledgement

We thank Edgar Roman-Rangel for co-inventing the HOOSC algorithm and providing the original code. We also thank Carlos Pallán Gayol, Guido Krempel, and Jakub Spotak for producing the data used in this paper. Finally, we thank the Swiss National Science Foundation (SNSF) and the German Research Foundation (DFG) for their support, through the MAAYA project.

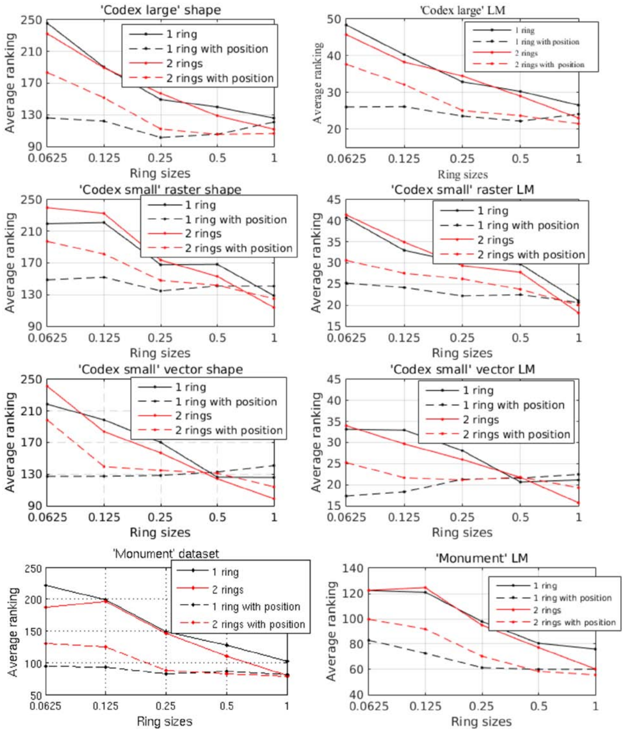


Figure 9. Retrieval results on each dataset, with various feature representation choices. (left) shape-base results; (right) incorporating glyph co-occurrence information.

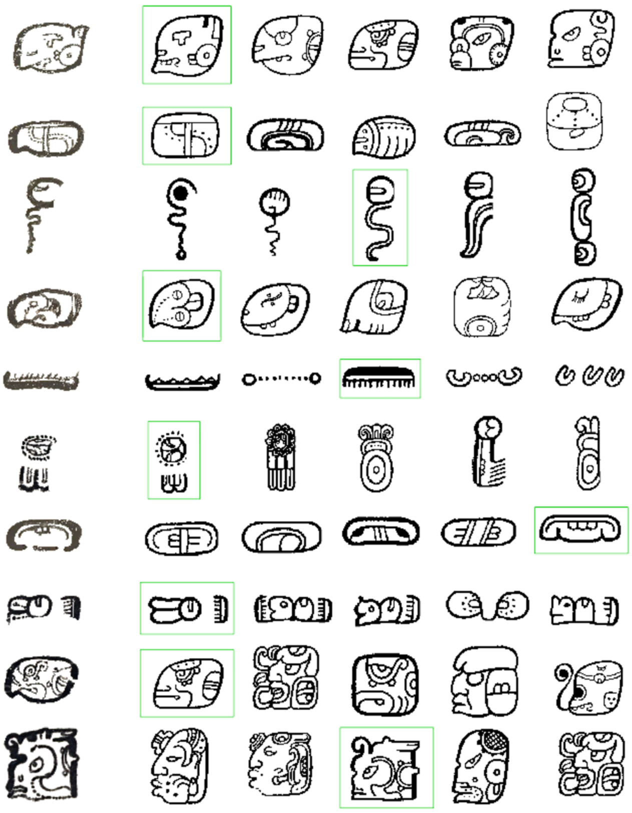


Figure 10. Example queries (first column) and their top returned retrieval results, ranked from left to right in each row. Groundtruth images are highlighted in green bounding boxes.

Bibliography

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape Matching and Object Recognition using Shape Contexts. *PAMI*. pp. 509-22.

Dalal, N., and Triggs, B. (2005). Histogram of Oriented Gradients for Human Detection. *In CVPR*. pp. 886-93.

Eitz, M., et al. (2012). Sketch-based shape retrieval. *ACM Transactions on Graphics*. **31**: 1-10.

Fischer, A., et al. (2012). The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries. *InterNational and InterDisciplinary Aspects of Scholarly Editing.*

Franken, M., and Gemert, J. C. (2013). Automatic Egyptian Hieroglyph Recognition by Retrieving Images as Texts. *ACM MM*, pp. 765-68.

Gatica-Perez, D., et al. (2014). The MAAYA Project: Multimedia Analysis and Access for Documentation and Decipherment of Maya Epigraphy. *Digital Humanities Conference.*

Hu, R., et al. (2015). Multimedia Analysis and Access of Ancient Maya Epigraphy: Tools to support scholars on Maya hieroglyphics. *IEEE Signal Processing Magazine*, pp. 75-84.

Roman-Rangel, E. (2012). *Statistical Shape Descriptors for Ancient Maya Hieroglyphs Analysis.* PhD thesis, École Polytechnique Fédérale de Lausanne.

Roman-Rangel, E., et al. (2011). Analyzing Ancient Maya Glyph Collections with Contextual Shape Descriptors. *IJCV*, pp. 101-17.

Stone, A.J. and Zender, M. (2011). *Reading Maya Art: A Hieroglyphic Guide to Ancient Maya Painting and Sculpture.* Thames and Hudson Limited Publisher.

Thompson, J. E. S. (1962). *A Catalog of Maya Hieroglyphs*. University of Oklahoma Press.

Vail, G., and Hernández, C. (2013). *The Maya Codices Database*, Version 4.1. A website and database available at http://www.mayacodices.org/.

Visualizing Mouvance: Towards an Alignment of Medieval Vernacular Text Traditions

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Institut für Informatik, Universität Leipzig, Leipzig, Germany

David Joseph Wrisley
dwrisley@gmail.com
American University of Beirut, Lebanon (Lebanese Republic)

Poetry in the Middle Ages changed as it was recopied, recited and passed along. *Mouvance* is a term used by late Swiss medievalist Paul Zumthor to designate the high degree of instability in medieval text traditions. Zumthor qualifies this instability as an "interplay between variant readings and reworkings," balancing both the textual, literary elements of written works with oral, performative ones (Zumthor, 1992; p.44). Twentieth-century text editing came to grips with this instability in large text traditions in different ways. For complex text traditions with a high degree of variance in medieval French, editors sometimes compiled one edition containing all the witnesses in successive chapters, as in the cases of *La vie de sainte Marie l'Egyptienne* (Dembowski, 1977) and *L'Evangile de Nicodème* (Ford, 1973) or they arranged them in a synoptic-style parallel edition as in the case of the *fabliaux* (Rychner, 1960). The alignment of such synoptic editions, although not discussed explicitly by the editors, was no doubt hand-ordered and its page layout was based on rough narrative equivalence of passages.

Cerquiglini has argued that, faced with such variance, the medievalist's "analysis must be comparative, not archeological" (Cerquiglini, 1999; p.44). Such parallel print editions provide readers with a visual frame for comparative reading of variant texts, inviting exploration and giving insight into processes of textual change. There are a number of problems in text traditions with oral influence. There are not only different kinds of variance (single word/string variance, half-line or hemistich variance, transposition and reorganization of rhyming verse lines or interpolation of entirely new lines), but patterns of variance are also not uniform across a text, making the desired comparative visualization of texts difficult. Alignments can also mix and confuse kinds of variance. What are ways of visualizing textual variation in a digital environment? This is the question our paper intends to explore.

Visualization strategies for historical text reuse vary according to the scale of the phenomenon and the nature of the texts involved (Franzini et al., 2015). Sophisticated visualizations for alignment exist at the micro-level, that is, at the level of the word such as the graph visualizations of TRAViz (Jänicke et al., 2015). They facilitate comparative readings of word-level variance in manuscript witnesses or translations. The text in such alignments is fully legible. A clean example of this can be found in the TRAViz alignments of English translations of the Bible (Figure 1), a textual use case in which units and sub-units of text are already commonly agreed upon by tradition (e.g. book, chapter, verse). On the other hand, solutions exist for macro-level text reuse, such as fingerprinting techniques (Jänicke and Geßner, 2015), creating distant patterns of textual similarity without showing the text (Figure 2).

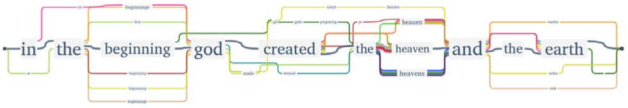


Figure 1: Micro-level alignment of 24 English translations of Genesis 1:1 (Jänicke et al., 2015)

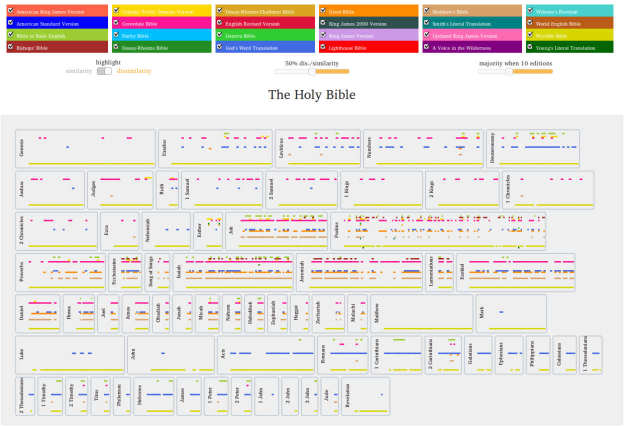
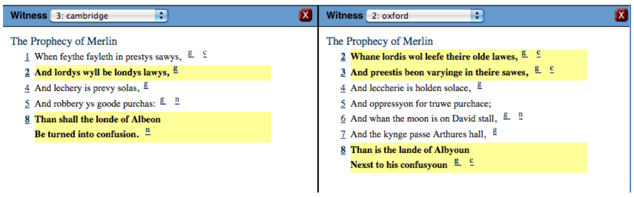


Figure 2: Macro-level fingerprint illustrates similarity between 24 English Bible translations (Jänicke and Geßner, 2015)

A critical discussion of the description and design of meso-level visualization of complex text traditions is missing. Not only are these text traditions large, but they are highly influenced by orality, that is, in Zumthor's words, they present a combination of textual and performative variance. Textual reworking at multiple scales (whole chunks of text, groups of lines, individual lines, sub-line strings) are challenges for both alignment and visualization. Our design attempts to translate the insights of Zumthor and Cerquiglini into an environment for visual exploration using two medieval text traditions: a tradition of short baudy tales known as the fabliaux , and in versions of the well-known epic the Chanson de Roland . These two traditions have been chosen for the kinds of complex variance they exhibit. Our paper focuses on so-called meso-level alignments that visualize patterns of textual variance higher than the word and verse line level, and stress both legibility and human interaction in visualizing patterns.

Existing methods for text alignment in digital environments, generally speaking, favor relatively stable texts with small variance. The Versioning Machine accepts texts encoded "according to TEI's Parallel Segmentation method" and "interprets the encoding, parsing out the text into its constituent parts" (Versioning Machine, 2015). The authors of the Versioning Machine offer a sample alignment of a medieval "Prophecy of Merlin" (Figure 3). The line-to-line alignment has been encoded by the textual scholar. Similar lines are visually connected using customary mouse behavior, however, variance within the line or across lines is not visualized.



Witness 3: cambridge

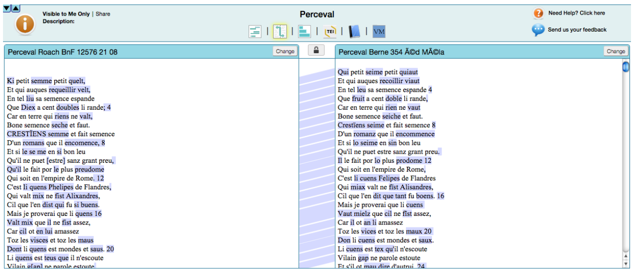
The Prophecy of Merlin
1 When feythe fayleth in prestys sawys, g c
2 **And lordys wyll be londys lawys,** g
4 And lechery is prevy solas, g
5 And robbery ys goode purchas: g n
8 **Than shall the londe of Albeon**
Be turned into confusion. n

Witness 2: oxford

The Prophecy of Merlin
2 **Whane lordis wol leefe theire olde lawes,** g c
3 **And preestis beon varyinge in theire sawes,** g c
4 And leccherie is holden solace, g
5 And oppressyon for truwe purchace;
6 And whan the moon is on David stall, g n
7 And the kynge passe Arthures hall, g
8 **Than is the lande of Albyoun**
Nexst to his confusyoun g c

Figure 3: Two versions of the Prophecy of Merlin visualized in Versioning Machine

Another environment for the collation of raw text and visualization of textual differences is JuxtaCommons (Wheeles and Jensen, 2013). The example in Figure 4 uses two editions of Chretien de Troyes' medieval romance *Perceval*, one based on BnF, ms. français 12576 (Roach, 1959) at left, and the other Bern, Burgerbibliothek ms. 354 (Méla, 1990) at right. There is basic alignment of the verse line and minor lexical or dialectal variance, and mouse over in JuxtaCommons allows basic comparative reading of word- and string-level variants. In their out-of-the-box implementation, both tools allow for easy comparison of two versions of the text, although the Versioning Machine has been implemented for larger comparison sets.



Perceval Roach BnF 12576 21 08

Ki petit semme petit quelt,
Et qui auques requeillir velt,
En tel liu sa semence espande
Que Diex a cent doubles li rande; 4
Car en terre qui riens ne valt,
Bone semence seche et faut.
CRESTÏENS semme et fait semence
D'un romans que il encomence, 8
Et si le se me en si bon leu
Qu'il ne puet [estre] sanz grant preu,
Qu'il le fait por le plus preudome
Qui soit en l'empire de Rome. 12
C'est li quens Phelipes de Flandres,
Qui valt mix ne fist Alixandres,
Cil que l'en dist qui fu si buens.
Mais je proverai que li quens 16
Valt mix que il ne fist assez,
Car cil ot en lui amassez
Toz les visces et toz les maus
Dont li quens est mondes et saus. 20
Li quens est teus que il n'escoute
Vilain g[ap] ne parole estoute

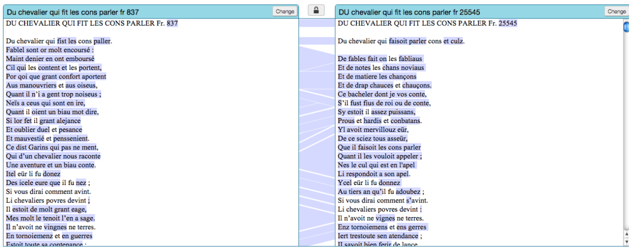
Perceval Berne 354 Ã©d MÃ©la

Qui petit seime petit quiaut
Et qui auques recoillir viaut
En tel leu sa semence espande 4
Que fruit a cent doble li rande,
Car en terre qui rien ne vaut
Bone semence seiche et faut.
Crestïens seime et fait semence 8
D'un romanz que il encommence
Et si lo seime en sin bon leu
Qu'il ne puet estre sanz grant preu.
Il le fait por lo plus prodome 12
Qui soit en l'empire de Rome,
C'est li cuens Felipes de Flandres
Qui miax valt ne fist Alisandres,
Cil que l'en dit que tant fu boens. 16
Mais je proverai que li cuens
Vaut mielz que cil ne fist assez,
Car il ot an li amassez
Toz les vices et toz les maux 20
Don li cuens est mondes et saux.
Li cuens est tex qu'il n'escoute
Vilain gap ne parole estoute
Et s'il ot mau dire d'autrui 24

Figure 4: Two editions of Chretien de Troyes' Perceval visualized in JuxtaCommons

When complex text traditions containing more than just variant readings, but also interpolations, extra or missing lines or a significant amount of orthographic variance are collated automatically in JuxtaCommons, their results, however, are nearly illegible. Figure 5 shows two witnesses of the same old French *fabliau*; the visual alignment, however, does an insufficient job at expressing the performative element of their *mouvance*.

Our design for a sufficient representation of variance in the *fabliaux* is shown in Figure 6. We use the intuitive quality of stream graphs (Byron et al., 2008) in order to support the analysis of aligned verses and to illustrate the transmission of one *fabliau* in four versions. The text editions are juxtaposed in columns in order to minimize edge crossings, in other words, we order the editions according to their similarity.



Du chevalier qui fit les cons parler fr 837

DU CHEVALIER QUI FIT LES CONS PARLER Fr. 837

Du chevalier qui fist les cons paller.
Fablel sont or molt encoursé :
Maint denier en ont emboursé
Cil qui les content et les portent,
Por qoi que grant confort aportent
Aus manouvriers et aus oiseus,
Quant il n'i a gent trop noiseus ;
Neïs a ceus qui sont en ire,
Quant il oient un biau mot dire,
Si lor fet il grant alejance
Et oublier duel et pesance
Et mauvestié et penssenient.
Ce dist Garins qui pas ne ment,
Qui d'un chevalier nous raconte
Une aventure et un biau conte.
Itel eür li fu donez
Des icele eure que il fu nez ;
Si vous dirai comment avint.
Li chevaliers povres devint ;
Il estoit de molt grant eage,
Mes molt le tenoit l'en a sage.
Il n'avoit ne vingnes ne terres.
En tornoiemenz et en guerres
Estoit toute sa contenance ;

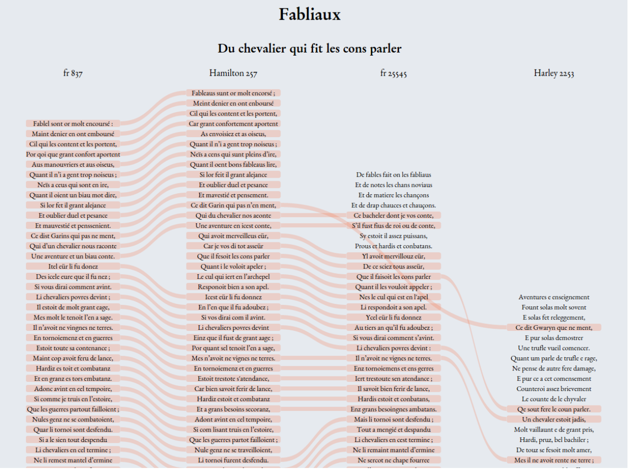
DU chevalier qui fit les cons parler fr 25545

DU CHEVALIER QUI FIT LES CONS PARLER Fr. 25545

Du chevalier qui faisoit parler cons et culz.

De fables fait on les fabliaus
Et de notes les chans noviaus
Et de matiere les chançons
Et de drap chauces et chauçons.
Ce bacheler dont je vos conte,
S'il fust fius de roi ou de conte,
Sy estoit il assez puissans,
Prous et hardis et conbatans.
Yl avoit mervillouz eür,
De ce sciez tous asseür,
Que il faisoit les cons parler
Quant il les vouloit appeler ;
Nes le cul qui est en l'apel
Li respondoit a son apel.
Ycel eür li fu donnez
Au tiers an qu'il fu adoubez ;
Si vous dirai comment s'avint.
Li chevaliers povres devint ;
Il n'avoit ne vignes ne terres.
Enz tornoiemens et ens gerres
Iert trestoute sen atendance ;
Il savoit bien ferir de lance

Figure 5: Two versions of a fabliau visualized in JuxtaCommons



Fabliaux

Du chevalier qui fit les cons parler

fr 837

Fablel sont or molt encoursé :
Maint denier en ont emboursé
Cil qui les content et les portent,
Por qoi que grant confort aportent
Aus manouvriers et aus oiseus,
Quant il n'i a gent trop noiseus ;
Neïs a ceus qui sont en ire,
Quant il oient un biau mot dire,
Si lor fet il grant alejance
Et oublier duel et pesance
Et mauvestié et penssenient.
Ce dist Garins qui pas ne ment,
Qui d'un chevalier nous raconte
Une aventure et un biau conte.
Itel eür li fu donez
Des icele eure que il fu nez ;
Si vous dirai comment avint.
Li chevaliers povres devint ;
Il estoit de molt grant eage,
Mes molt le tenoit l'en a sage.
Il n'avoit ne vingnes ne terres.
En tornoiemenz et en guerres
Estoit toute sa contenance ;
Maint cop avoit feru de lance,
Hardiz es toit et combatanz
Et en granz es tors embatanz.
Adonc avint en cel tempoire,
Si comme je truis en l'estoire,
Que les guerres partout failloient ;
Nules genz ne se combatoient,
Quar li tornoi sont desfendu.
Si a le sien tout despendu
Li chevaliers en cel termine ;
Ne li remest mantel d'ermine

Hamilton 257

Fableaus sunt or molt encorsé ;
Meint denier en ont enboursé
Cil qui les content et les portent,
Car grant confortement aportent
As envoisiez et as oiseus,
Quant il n'i a gent trop noiseus ;
Neïs a cens qui sunt pleins d'ire,
Quant il oent bons fableaus lire,
Si lor feit il grant alejance
Et oublier duel et pesance
Et mavestié et pensement.
Ce dit Garin qui pas n'en ment,
Qui du chevalier nos aconte
Une aventure en icest conte,
Qui avoit merveilleus eür,
Car je vos di tot asseür
Que il fesoit les cons parler
Quant i le voloit apeler ;
Le cul qui iert en l'archepel
Responoit bien a son apel.
Icest eür li fu donnez
En l'en que il fu adoubez ;
Si vos dirai com il avint.
Li chevaliers povres devint
Einz que il fust de grant aage ;
Por quant sel tenoit l'en a sage,
Mes n'avoit ne vignes ne terres.
En tornoiemenz et en guerres
Estoit trestote s'atendance,
Car bien savoit ferir de lance,
Hardiz estoit et combatanz
Et a grans besoins secoranz,
Adont avint en cel tempoire,
Si com lisant truis en l'estoire,
Que les guerres partot failloient ;
Nule genz ne se travailloient,
Li tornoi furent desfendu.

fr 25545

De fables fait on les fabliaus
Et de notes les chans noviaus
Et de matiere les chançons
Et de drap chauces et chauçons.
Ce bacheler dont je vos conte,
S'il fust fius de roi ou de conte,
Sy estoit il assez puissans,
Prous et hardis et conbatans.
Yl avoit mervillouz eür,
De ce sciez tous asseür,
Que il faisoit les cons parler
Quant il les vouloit appeler ;
Nes le cul qui est en l'apel
Li respondoit a son apel.
Ycel eür li fu donnez
Au tiers an qu'il fu adoubez ;
Si vous dirai comment s'avint.
Li chevaliers povres devint :
Il n'avoit ne vignes ne terres.
Enz tornoiemens et ens gerres
Iert trestoute sen atendance ;
Il savoit bien ferir de lance,
Hardis estoit et conbatans,
Enz grans besoingnes ambatans.
Mais li tornoi sont desfendu ;
Tout a mengié et despandu
Li chevaliers en cest termine ;
Ne li remaint mantel d'ermine
Ne sercot ne chape fourree

Harley 2253

Aventures e enseignement
Fount solas molt sovent
E solas fet releggement,
Ce dit Gwaryn que ne ment,
E pur solas demostrer
Une trufle vueil comencer.
Quant um parle de trufle e rage,
Ne pense de autre fere damage,
E pur ce a cet comensement
Counteroi assez brievement
Le counte de le chyvaler
Qe sout fere le coun parler.
Un chevaler estoit jadis,
Molt vaillaunt e de grant pris,
Hardi, pruz, bel bachiler ;
De touz se fesoit molt amer,
Mes il ne avoit rente ne terre ;

Figure 6: Four versions of a fabliau in our design

The visualization is available at http://informatik.uni-leipzig.de:8080/Fabliaux/. Clicking on a specific verse line produces a TRAViz micro-view of the line-level variance (Jänicke et al., 2015), whereas the larger meso-view of this portion of the *fabliau* allows patterns between and across verse lines to be clearly ascertained. V ariance in this genre maintains prosody and avoids hypermeter; *mouvance* is characterized here by the interpolation of larger narrative multi-line fragments of text. The exception to this general rule is manuscript Harley 2253 in the column at far right where the narrative is reconstituted almost completely around sparse line re-use, illustrating what Rychner calls in the subtitle of his book "deteriorization" [dégradation] (Rychner, 1960). *Mouvance* occurs in multi-line "chunks," visible in highly legible streams of text re-use. This provides much more insight into textual transformation than reading Rychner's hand-aligned synoptic edition, with the streams here drawing our attention to patterns of re-used text. In his edition, blank spaces and horizontal lines in the page layout effect a more uncertain alignment.

Another example of what we are calling meso-level alignment is the visualization of one stanza from the *Chanson de Roland* tradition contained in six manuscripts.

Figure 7 illustrates laisse 1 of the Lavergne fragment, absent from Oxford, Bodleian Library manuscript Digby, 23. An interactive version of the alignment is available at http://informatik.uni-leipzig.de:8080/roland/index2.html?ftsize=11 .

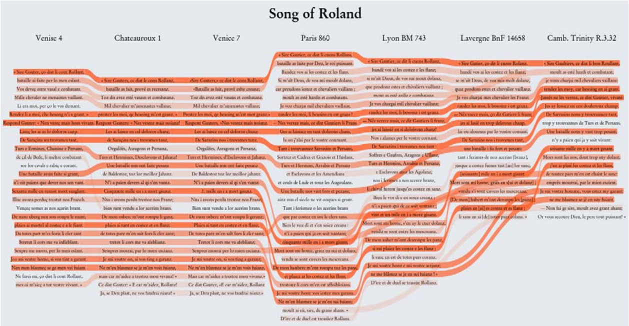


Figure 7: Six editions and one fragment of the Chanson de Roland in our design

To indicate how often lines recur across the whole manuscript tradition, we use streams colored with varying saturation. Highly saturated colors indicate frequently repeated passages, whereas less saturated colors indicate less repeated ones. Such a feature allows for a “consensus” visualization of the tradition at this meso scale. It is easy to see the more complex, transpositional variance of lines in the Chanson de Roland . This compositional feature of French epic visualized here needs to be studied across the entire corpus of seven manuscripts and three fragments. It is perhaps on this point, however, that a principle of visualization clashes with the visual expectations of a medieval textual scholar. As in the above example, we ordered the editions to reduce the number of crossing streams and to maximize legibility. This is potentially at odds with readers who expect to see temporality of manuscript dating represented in the visualization. More thinking needs to be done about the visual semantics of such a large tradition. In such a case, a perfect order that produces less clutter might be hard to determine, and we will be required to extend our proposed design.

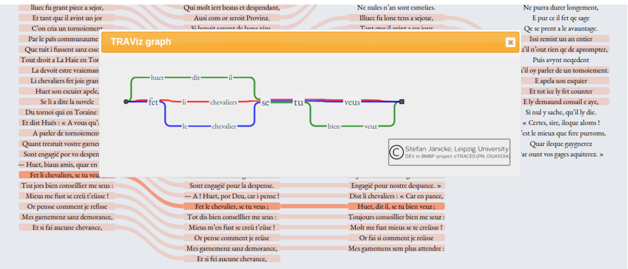

TRAViz graph
huet
dit
il
fet
li
chevaliers
le
chevalier
se
tu
veus
bien
veuz
Stefan Jänicke, Leipzig University
DEV in BMBF-project eTRACES (PN: 01UA1101A)
Fet le chevalier, se tu veus ;
Huet, dit il, se tu bien veuz ;

Figure 8: Visualization of an aligned verse line in three versions of a fabliau

We began with the alignment of the fabliaux and Chanson de Roland since they are traditions where editors have indexed alignment manually, either using page layout or by numerical cross referencing of the stanzas . As the example shown in Figure 8 illustrates, lines are not identical word for word, but rather our alignment has asserted basic diegetical equivalence.

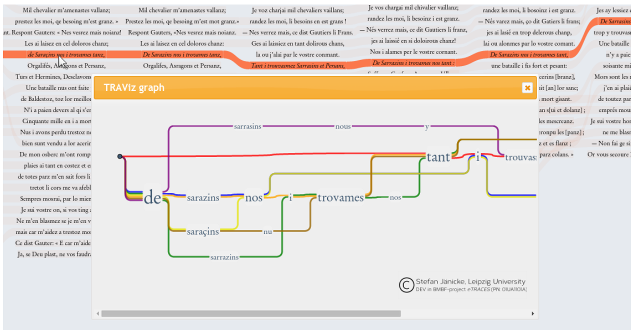

TRAViz graph
de
sarrasins
nous
y
sarazins
nos
i
trovames
tant
i
trouvas
saraçins
nu
sarrazins
nos
Stefan Jänicke, Leipzig University
DEV in BMBF-project eTRACES (PN: 01UA1101A)
De Sarazins nos i trovames tanz,
Tant i trouvasmez Sarrasins et Persans,
De Sarrazins i trovames nos tant :
De Sarrasins nous y trouvames tant,

Figure 9: Alignment of an aligned verse among seven Chanson de Roland editions

In the example portrayed in Figure 9, performative mouvance is reflected in the recombination of words in the line.

In developing our research on this topic, we intend to pursue more granular alignment using computational means, down to the line and perhaps the hemistich. Based on the Relative Edit Distance of strings (Jänicke et al., 2015), we aim to determine spelling variants and rhyme sets automatically. Combined with an n-gram analysis, we will support the discovery and alignment of similar text passages. Due to the high degree of orthographic variance of the medieval French language, a purely computational approach might lead to a high number of false positive alignments. Therefore, we will design a visual analytics system that includes the human in the alignment process in order to configure appropriate settings that maximize the alignment of re-used text passages. This visual analytics process includes the supervised training of a classifier that supplies the parameters steering the alignment. Iteratively, computationally aligned text fragments are scored by the textual scholar according to their relevance. After each scoring session, the alignment will be recomputed taking the scholar's justifications into account.

In contrast to the manual alignment illustrated in the visualizations above (Figures 6-9), such a semi-automated process will potentially yield a very different picture in particular with respect to the degree of difference the human allows in string matching and the presence of shared n-grams required to align certain text passages. A further opportunity of this generic visual analytics approach is its straightforward adaptability to editions in other languages than medieval French. With such a user-centered approach, alignment is not to be understood as a final product, but rather a process, for understanding variant text traditions, exploring their intricacies and supporting generation of hypotheses about textual behavior.

Bibliography

Anonymous (2005). La Chanson de Roland / The Song of Roland: The French Corpus . Ed. J. Duggan. Geneva: Droz.

Byron, L., and Wattenberg, M. (2008). Stacked graphs–geometry & aesthetics. In Visualization and Computer Graphics, IEEE Transactions. 14.6 (2008): 1245-1252.

Cerquiglini, B. (1999). In Praise of the Variant: A Critical History of Philology . Baltimore/London: The Johns Hopkins Press.

Dembowski, P. (1977). La vie de sainte Marie l'Egyptienne, versions en ancien et en moyen français . Geneva: Droz.

Ford, A.E. (1973). L'Evangile de Nicodème: les versions courtes en ancien français et en prose . Geneva: Droz.

Franzini, G., Franzini, E., Büchler, M. (2015). Historical Text Reuse: What Is It?. Available at: http://etrap.gcdh.de/?page_id=332 [accessed 3 March 2016].

Jänicke, S. and Geßner, A. (2015). A Distant Reading Visualization for Variant Graphs. In Proceedings of the Digital Humanities 2015 .

Jänicke, S., Geßner, A., Franzini, G., Terras, M., Mahony, S. and Scheuermann, G. (2015). TRAViz: A Visualization for Variant Graphs. In Digital Scholarship in the Humanities , 30(suppl.1): i83–i99. Available at: http://www.informatik.uni-leipzig.de/~stjaenicke/TRAViz.pdf [accessed 27 October 2015]

Méla, C. (1990). Le Roman de Perceval. Édition du ms. 354 de Berne, traduction critique, présentation et notes . By Chrétien de Troyes. Paris, Librairie générale française.

Moffat, M. (2014). The Châteauroux Version of the "Chanson de Roland": A Fully Annotated Critical Text . Berlin/Boston: De Gruyter.

Roach, W. (1959). Le roman de Perceval ou le Conte du Graal publié d'après le ms. fr. 12576 de la Bibliothèque nationale . By Chrétien de Troyes. Geneva, Droz.

Rychner, J. (1960). Contribution à l'étude des fabliaux: variantes, remaniement, dégradations . Geneva: Droz.

Versioning Machine. (2015). http://v-machine.org/samples/prophecy_of_merlin.html [accessed 27 October 2015]

Wheeles, D. and Jensen, K. (2013). Juxta Commons. In Proceedings of the Digital Humanities 2013 .

Zumthor, P. (1992). Toward a Medieval Poetics . Minneapolis: University of Minnesota Press.

Topic Modeling Literary Quality

Kim Jautze
kim.jautze@huygens.knaw.nl
Huygens ING, Royal Netherlands Academy of Arts and Sciences, Netherlands

Andreas van Cranenburgh
andreas.van.cranenburgh@huygens.knaw.nl
Huygens ING, Royal Netherlands Academy of Arts and Sciences, Netherlands; Institute for Logic, Language and Computation, University of Amsterdam

Corina Koolen
c.w.koolen@uva.nl
Institute for Logic, Language and Computation, University of Amsterdam

Introduction

To what extent can topic models explain variation in perceptions of literary quality? We try to find correlations between topics and judgments of literary quality using a topic model of 401 recent bestselling Dutch novels. Instead of examining topics on a macro-scale in a geographical or historical interpretation (e.g., Jockers 2013; Riddell 2014), we take a new perspective: whether novels have a dominant topic in their topic distributions (mono-topicality), and whether certain topics may express an explicit or implicit genre in the corpus. We hypothesize that there is a relationship between these aspects of the topic distributions and perceptions of literary quality. We then interpret the model by taking a closer look at the topics in a selection of the novels.

Riddle survey and corpus

This research is part of a Dutch computational humanities project called The Riddle of Literary Quality. In the project we aim to identify textual features that may play a role in readers' evaluations of a novel as being good or bad and as high or low literature. We analyze a corpus of 401 contemporary Dutch-language (including translated) novels in search of textual features they have in common. Within our corpus there is a small variety of novelistic genres, which can be roughly divided into suspense, romantic and general novels. The readers' judgments were gathered in a large online survey. We asked a general public to rate the novels they had read on a 7-point scale from *definitely not* through *highly* literary. Approximately 14,000 respondents participated, providing us with much data on the perceived quality of our 401 novels.[1] The mean rating over all 401 novels is 4.2, with 2.1 being the lowest rating for *Fifty Shades of Grey* by E.L. James, and 6.6 the highest for Julian Barnes' *The Sense of an Ending*.

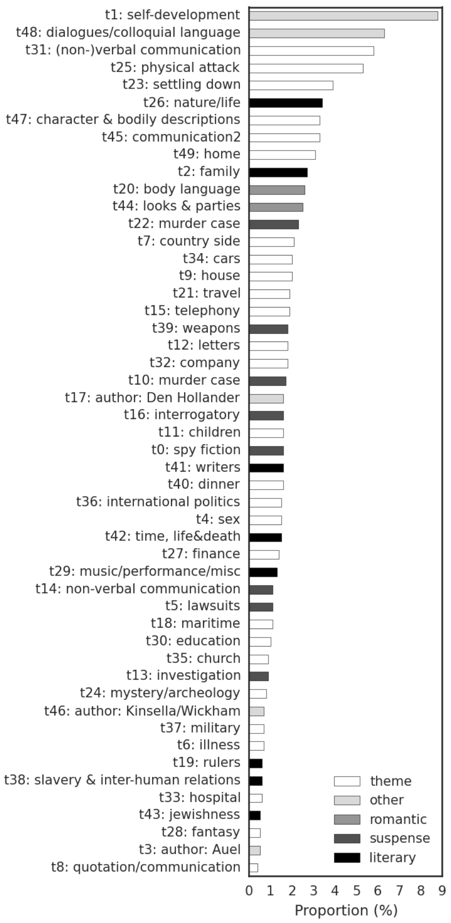


Figure 1: Overview of topics, sorted by proportion of the corpus

Topic model

A topic model aims to automatically discover topics in a collection of documents. We use Latent Dirichlet Allocation (Blei et al., 2003), which assumes the documents have been generated from a fixed number of probability distributions (the topics) over words. The topics reflect word co-occurrence patterns. We preprocess the novels by lemmatizing the words, removing punctuation, function words and names, and splitting the remaining text in chunks of 1000 tokens. We use MALLET to create a topic model with 50 topics. Fig. 1 shows an overview of the topics with their proportion across the corpus.

We have attempted to identify topics for novels with high literary ratings, and topics specific for suspense and romantic novels. According to Jockers and Mimno (2013), the topics can be used to identify literary themes. They use the terms "theme" and "topic" as "proxies for [...] a type of literary content that is semantically unified and recurs with some degree of frequency or regularity throughout and across a corpus" (p. 751). We found that three topics are specific to a single author (for instance t3), and about a third seem genre specific. By inspecting the most important words for each topic we found that most topics (genre related or not) indeed cohere with certain themes (cf. Fig.1). This suggests that the choice for 50 topics is neither too small nor too high.

Quantitative analysis

We aim to gain insight into the distribution of topics in relation to the literary ratings of the novels (predicting literary ratings is not the main aim here). In order to interpret the topic distributions, we introduce the concept of mono-topicality.[2] A mono-topical novel contains little diversity in topic distribution, which means that one or two topics are dominant throughout the novel. A novel which shows more variation in topics has a more even distribution of topics, i.e., such a novel has a larger topic diversity. Fig. 2 shows an example of both cases.

The x-axis shows the distribution of topics, sorted from least to most prevalent. In John Grisham's *The Appeal*, topic 5 ("lawsuits") has a proportion of 47.8 % of all 50 topics. This novel is more mono-topical than the Franzen's *Corrections*, which has a more balanced distribution of topic proportions.

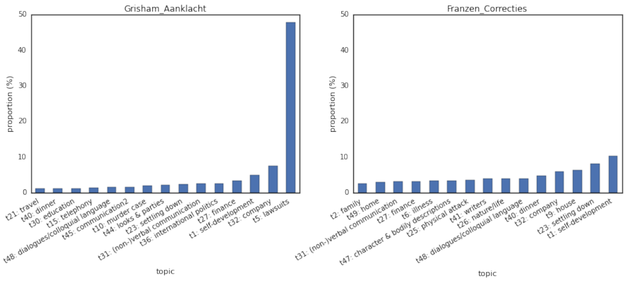


Figure 2: Distribution of the top 15 topics in novels with high (left) and low (right) mono-topicality

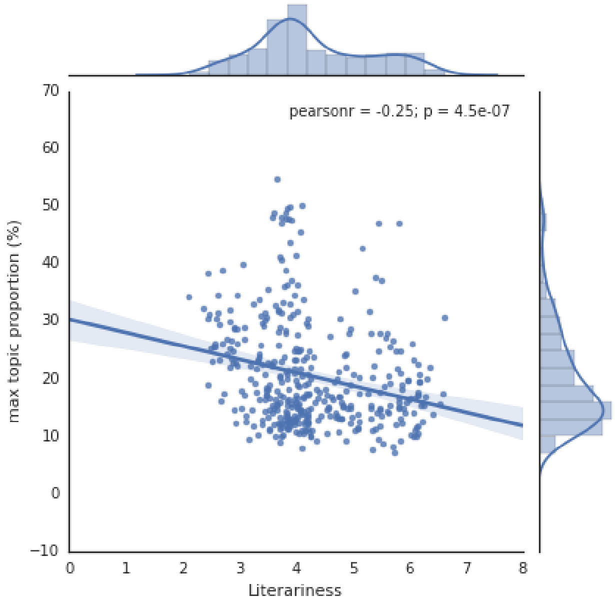


Figure 3: Correlation between share of themost prominent topic per book and mean literariness ratings

Author_Title	Survey rating	% topic 29	Relationships			Complications		Artistic profession		True story	
			Friend / family	Parent-child	love	rivalry	illness & Heath	writer / editor	Rother	autobiographic	history
Hart_Verlovingstijd	5.9	24.8	X			X					
Rosenboom_ZoeteMond	6.2	22.5			X	X					
Lanoye_Sprakeloos	6.4	21.8		X			X	X		X	
Dewulf_KleineDagen	6.0	19.6	X					X		X	
Rosenboom_Mechanica	6.2	1 7.0						X			
Heijden_Tonio	6.3	16.8		X			X	X		X	
Verhulst_LaatsteLiefdeVan	5.8	16.5		X		X					
Lanoye_HeldereHemel	5.8	15.7	X			X					X
Springer_Quadriga	6.0	15.7			X			X			X
Mortier_GestameldLiedboek	6.5	15.5		X			X	X		X	
Kooten_Verrekijker	5.0	15.2		X				X		X	
Moor_SchilderE	5.9	14.7			X				X		X
Meisje											
Zwagerman_Duel	5.5	14.6							X		
Giphart_IJsland	5.3	13.0		X					X		
Dorrestein_Stiefmoeder	5.5	8.7	X	X							

Table 1: Themes in fifteen highly literary novels, all of which are originally Dutch Table 2: Six topics from the model that address themes present in the fifteen highly literary novels, but which are not the most prominent as topics in those novels

We hypothesize that the less mono-topical a novel is, the higher the perceived literariness by readers will be. And indeed, Fig. 3 shows that there is a statistically significant correlation between the diversity of topics of a book and its perceived literariness. Books with a single, highly prominent topic, such as Grisham's, tend to be seen as less literary.

Interpretation

There are several possible explanations for the correlation. Genre novels could have a tendency to single out certain topics, as they deal with more 'typical' or genre-specific subject matter than do general novels. If this were the case, we would simply be finding that genre novels are considered to be less literary than general novels, and this would tell us little about literary quality in a more general sense. General novels in the other hand, deal with all sorts of subjects and themes, across and beyond 'genre' subjects, and therefore a topic model may not be able to single out thematic groups of words common to these novels, and thus may not find one single prominent topic. A third explanation could be that highly literary novels *do* deal with specific themes and subjects which are also part of genre novels, but that these are described in wordings that are more implicit or subtle, and therefore do not come up as single, clear topics. If this were the case, that would bring us closer to an explanation of what topics have to do with literary quality. These explanations are not mutually exclusive and we will explore the topic model here to examine the value of the second and third explanation.

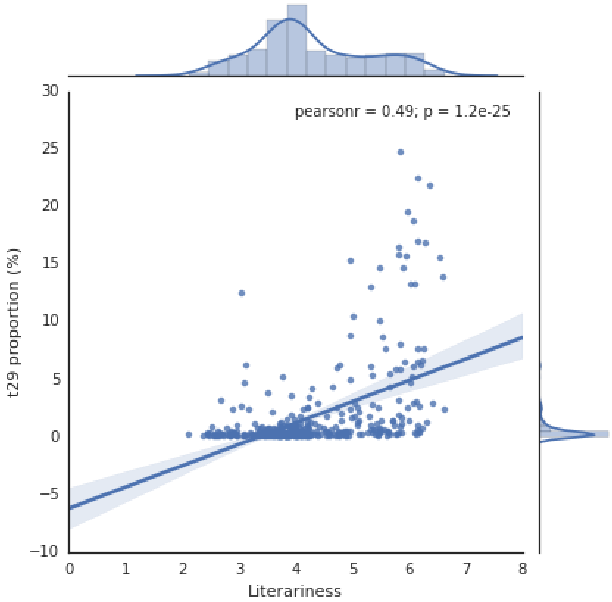


Figure 4: Correlation between topic 29 proportion and mean literariness ratings

The topic that shows the highest correlation (r=0.49) with literary appreciation is topic 29; cf. Fig. 4. This topic is most prominent in fifteen originally Dutch general novels. The twenty words in topic 29 with the highest weights are *begin, music, play, occasion, first, the first, sing, only, year, one, stay, sometimes, even, new, own, always, high, exact(ly), bike, appear*. They show little coherence, making it hard to interpret their context, although 'music and performance' appears to be part of it. To find out more about the novels in which this topic is prominent, we consult a Dutch website maintained by librarians called *Literatuurplein*, which provides information on the themes and content of Dutch novels.

Most of these novels show similarities in themes, such as family relationships. In ten of the novels the protagonist has an artistic profession: a couple of writers, a painter and a stand-up comedian. None of them has a musical or acting career, despite the 'music and performance' words; and vice versa, none of the twenty most prominent words concern writing.

All in all, at first glance topic 29 seems not to address the themes and content of the novels, whereas most other topics in the model do concern specific themes (cf. Fig. 1 and Table 2).

Topic	Name	Top 10 words with highest weight
2	Family relations I	father, mother, child, year, son, girl, brother, woman, older, daughter
6	Health I	doctor, body, pain, illness, pill, blood, death, medicine, child, patient
11	Family relations II	child, mother, mom, baby, dad, little, cry, hand, time, grandma
12	Writing & memories	picture, letter, write, read, paper, book, year, day, enveloppe, memories
33	Health II	doctor, hospital, patient, women, bed, lie, nurse, room, hall, hour
41	Novels	book, writing, story, year, word, writer, human, novel, time

Table 2: Six topics from the model that address themes present in the fifteen highly literary novels, but which are not the most prominent as topics in those novels

For instance, topic 2 and 11 address family relations, topic 12 and 41 are about writing novels, and topic 6 and 33 concern health issues. These topics are present, but as smaller topics. This shows that the second explanation, of the general novels not sharing themes, is not valid. It could be an indication though that the highly literary novels indeed use a more subtle way of describing themes similar to other novels in our corpus, our third explanation. As a final note, in topic 29 there are proportionally more adverbs than in the other topics mentioned, which contain more nouns. Perhaps this shows that style is a more shared element in literary novels than the choice of words. In other words, this brief analysis shows that there is merit to our third explanation. This will therefore become a new hypothesis for further research.

Conclusion

We have explored a topic model of contemporary novels in relation to genre and literariness, and shown that topic diversity correlates with literary ratings. Most topics express a clear theme or genre. However, topic 29, the most literary topic, does not. It rather appears to be associated with a particular Dutch literary writing style.

Bibliography

Algee-Hewitt, M., Heuser R., and Moretti, F. (2015). On paragraphs. Scale, themes, and narrative form. Stanford Literary

Lab pamphlet 10. http://litlab.stanford.edu/LiteraryLabPamphlet10.pdf

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jockers, M. L. and Mimno, D. (2013). Significant Themes in 19th-Century Literature. *Poetics* **41**(6):750–69. http://dx.doi.org/10.1016/j.poetic.2013.08.005

Jockers, M. L (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Riddell, Allen (2014). How to read 22,198 journal articles: Studying the history of German studies with topic models. In Erlin, M. and Tatlock, L. (eds), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Rochester, New York: Camden House, pp. 91–114.

Notes

[1] Extensive details on the survey will be published in two articles, one of which is in submission.

[2] After submitting this abstract, we discovered the Literary Lab pamphlet by Algee-Hewitt et al. (2015), who independently devised a concept called mono-topicality.

Teaching Digital Humanities Through a Community-Engaged, Team-Based Pedagogy

Andrew Jewell
ajewell2@unl.edu
University of Nebraska-Lincoln, United States of America

Elizabeth Lorang
llorang2@unl.edu
University of Nebraska-Lincoln, United States of America

Based in five years of experience teaching digital humanities (DH) students at the University of Nebraska-Lincoln (UNL), our paper tells the story of an evolving approach to teaching "practical" skills. We argue that the most important skills DH students need to learn are not particular programming languages or specific research methodologies, but team-based problem-solving. Furthermore, an effective way to achieve this learning is for students to work together to design and build a digital project that addresses a real challenge, draws upon their commitment to the humanities, and serves the mission of a local organization.

Background

When UNL formed its Graduate Certificate Program in Digital Humanities, the organizing committee knew it wanted students to have a substantial engagement with the intellectual theory undergirding digital humanities as well as with DH praxis. To accomplish this, UNL created two courses to form the heart of the Certificate: the Interdisciplinary Reading Seminar in Digital Humanities and the Internship in Digital Humanities.

The Internship in Digital Humanities, originally available only to graduate students, embedded students within faculty DH projects at UNL. Students worked on these projects for seven hours per week and spent one hour in class learning some "basic skills" for digital humanists. After running the course this way for two years, we determined we were not fulfilling the goals of the curriculum or the needs of the students. In the best cases, students were fully integrated into project teams and were challenged with new experiences. Other students, however, performed menial and repetitive work throughout the semester. We also underestimated the challenge of making students collaborators in this limited time, especially when project staff faced deadlines and needed to focus on production rather than instruction. The weekly class sessions, too, were mere introductions; students could not truly learn new skills because their projects did not provide opportunities to work with the introduced technologies and strategies.

The mediocrity of this approach and the inclusion of the course in a new Undergraduate Minor in Digital Humanities forced us to think more deeply about what it means to teach DH project development. We wanted a higher level of student investment in the work and for students to be involved from conception to implementation. And opening the course to undergraduate students made us even more aware of the need to present students with varied projects, not just those emerging from faculty members at a research-intensive university.

Immersive, Community-Based Approach to Digital Humanities Teaching and Learning

For the revised version of the course, known as the Digital Humanities Practicum and launched in 2014, we adopted a service learning--or community-based learning--model. Both undergraduate and graduate students enroll in the course, and it is cross-listed among several departments including Anthropology, English, History, and Modern Languages & Literatures. (We are faculty members in the UNL Libraries, and the Libraries also play an important role in the course.) Under the new design, we partner with local organizations who have identified challenges suited to technological, humanities-engaged solutions. Over the course of the semester, students respond to those challenges, first conceptualizing a solution, iteratively building their solution (and sharing iterations with classmates), and then presenting their solution to a public audience at the end of the semester. The practicum engages and implements key values of community-based learning, including a "recursive style; direct, high-impact method; and emphasis on abstraction embedded in practice" (Grobman and Rosenberg, 2015). In addition, the course now advances a team-based experience that focuses not only on academia but looks outward to the humanities' roles in society more broadly.

A fundamental difference between the Digital Humanities Practicum and the earlier internship course is that the Practicum focuses on team-based problem-solving rather than specific technical skills. For example, Humanities Nebraska, a state-wide humanities advocacy organization, challenged the students to improve communication about their annual Chautauqua event while engaging new audiences. In response, a team of undergraduate students proposed a mobile application that would serve as an information platform and provide opportunities for social media engagement. Entering the course, the students had limited experience with web technologies and no experience with mobile application development. While they researched what might be involved in creating a mobile application, we reached out to others on the UNL campus who could work with students to help them learn specific skills and made sure they would have access to necessary hardware (such as a variety of mobile devices for testing) and software. By semester's end, the application was available in the Google Play store, and shortly after was published to the Apple App Store. During this experience, the students learned much more than new technology proficiencies. They performed research about Chautauqua and the Chautauqua theme ("Free Land"), considered how best to communicate this information to the audiences they sought to reach, and interacted effectively with their client and mentors about their ideas--including accepting and responding to criticism of approaches that were not working.

This model requires significant flexibility on the part of the instructors and students. The syllabus is largely unfixed, as it must respond to the students and their needs, based on their background and experiences and also on the solutions they seek to pursue. Therefore, most of the fifteen-week semester cannot be planned more than a week or two in advance. Students, however, use an agile development approach so that they and we learn early where they will confront difficulties in implementing their solutions and what resources--whether people, hardware, software, or strategies--we need to connect them with in order for them to develop their projects.

Successful implementation of this model also requires that faculty and students are frank about knowledge limitations. As the instructors, we confess at the beginning of the course that we don't ourselves know everything the students will need to learn to be successful. But what we offer the students--and model for them--is the ability to figure out the necessary skills and seek appropriate resources. As we routinely tell the students, the requirements to solve these problems are not technical skills, but courage and perseverance. Our students have learned to weather discomfort not only because we implore them to do so, but because the iterative development model insists upon it. By having the students produce and demonstrate results early and often, we get them accustomed to a new kind of relationship with their coursework and to problem solving.

Based on our observations and student evaluations, it is this iterative process that imparts the learning. Furthermore, student investment in the projects is encouraged by the "realness" of the challenges. Unlike many classroom assignments, the problems in the Digital Humanities Practicum are authentic challenges brought in by external organizations with real missions, and their work can have application beyond the classroom. Students have assisted the Nebraska Commission on Indian Affairs, for example, in their effort to secure federal legislation designating an historic trail. For the next iteration of the course, offered in Spring 2016, we plan to work with organizations that are not principally humanities organizations. These include a children's museum, a community supported agriculture and food education organization, and a social justice organization. Our goal is to broaden understanding of where humanities work can happen as well as demonstrate possibilities for solving problems by joining diverse areas of expertise. While we do not yet know the outcomes of the 2016 Practicum, the course will conclude before DH2016, and we will share both the student projects as well as an evaluation of the approach.

Bibliography

Grobman, L. and Rosenberg, R. (2015). Introduction: Literary Studies, Service Learning and the Public Humanities. *Service Learning and Literary Studies in English*, ed. Laurie Grobman and Roberta Rosenberg (New York: MLA), pp. 1-39.

Jakacki, D. and Faull, K. (2014). Digital Learning in an Undergraduate Context: Promoting Long Term Student-Faculty (and Community) Collaboration in the Susquehanna Valley, PA. In: *Digital Humanities 2014*. Lausanne, Switzerland, July 2014. http://dharchive.org/paper/DH2014/Paper-492.xml (accessed 2 March 2016).

Rockwell, G. and Sinclair, S. (2012). Acculturation and the Digital Humanities Community. In Hirsch, B. (ed), *Digital Humanities Pedagogy: Practices, Principles, and Politics*. UK: Open Book Publishers. http://www.openbookpublishers.com/htmlreader/DHP/chap07.html#ch07 (accessed 2 March 2016).

Smith, D. (2014). Advocating for a Digital Humanities Curriculum: Design and Implementation. In: *Digital Humanities 2014*, Lausanne, Switzerland, July 2014. http://dharchive.org/paper/DH2014/Paper-665.xml (accessed 2 March 2016).

Notes

[1] Full details of the UNL Graduate Certificate Program are available at http://www.unl.edu/dhcert/

[2] Our approach acknowledges Jakacki and Faull's perspec-

tive that undergraduate digital humanities education often focuses on specific tools or techniques rather than "habits of mind" (2014). Our course and pedagogy also take up the "three major questions in digital humanities pedagogy" raised by Smith: is there a "common core of learning objectives" around which DH programs should be structured? How vocational should DH curricula be? Should DH curricula focus on skills, method, "or critical perspectives on technology and its application"? (2014).

[3] The skills we teach with this class are consistent with those outlined by Rockwell and Sinclair (2012): working in interdisciplinary teams, managing projects, applying digital practices, and explaining technologies.

Exploring and Evaluating Cartographic (miss)Representation in a Sample of Web-based Geohumanities Projects

Catherine Emma Jones
catherine.jones@cvce.eu
CVCE, Luxembourg

Introduction

Given the rise in recent years of GeoHumanities projects, this paper considers how historical data are translated and represented geographically within online mapping projects, giving insights into how these representations influence acquired meaning. GeoHumanities predominantly use geographical tools to create, present and explore different types of historical evidence and resources, for example original paper maps transformed into digital representations or geocoding of place names within texts. Factors impacting the resulting representation of historical evidence in geographical form include: choice of the underlying geographic data models, form and style of the original (historical) objects, as well as the selection, transformation and encoding practices. In this study we observe which data models are used and their context, to understand how these choices may influence acquired meaning and cognitive understanding.

Method

The GeoHumanities special interest group of the Alliance for Digital Humanities Organisations has a catalog of 312 geohumanities based projects (data from July 2014). From this list a sample of 30 projects were selected (sample selected based on language, availability and production since 2010) . In these projects, the cartographic representation of historical evidence was evaluated according to a set of 70 different criteria grouped into 4 categories: (1) data structures used to model data; (2) forms of representation (typology, symbolism, use of transparency and uncertainty); (3) data interaction possibilities offered to users and (4) spatial data analysis possibilities offered to users.

The following research questions were investigated:

- How are historical data represented in web mapping applications?
- How rich are geohumanities interfaces?
- How do they influence knowledge acquisition?

In this paper we consider the first research question.

Results: Exploring raster representations

In the sample we observed a mix of both vector and raster data models. Of the 30 projects, 40% (n=12) used raster data models to integrate georeferenced versions of historical paper maps, with 4 types of use scenarios (see figure 1). Firstly, either paper map sheets belonging to a map series were stitched together or individual maps were used as a basemap layer enabling viewing of historical context or comparison of raster map to digitally encoded vector data. The digital versions of the paper map were presented as an optional data layer in the interface that, in the main, could be switched on or off by the user. For 26 out of 30, the historical map was overlaid on top of a contemporary raster base map which displayed either a street map or an aerial view. In only 2 very early period maps, the Agas Map of Early Modern London and The Gough Map of Great Britain, was it not possible to view a contemporary reference basemap.

The second and third types of historical paper map representations in raster format were observed in crowdsourcing methodologies. In these projects users are asked to interact with the map to create new digital data. One category of projects asked members of the public to undertake the transformation from a digital scan of the paper map to a true digital raster using the processes known as georeferencing. The other made use of georeferenced historical maps and asked users to vectorise historical data. Users were asked to encode labels or draw and confirm shapes of buildings. In this type of task the users are effectively tracing the historical map to create a vector representation. The final use observed focused on the search of historical data through the map interface to find metadata records and locate historical maps. The coordinate extents of the original paper maps were used to draw a minimum bounding rectangle within the web-map interface to highlight the represented areas. Users could then search for places and discover which old maps are available digitally at different archives.

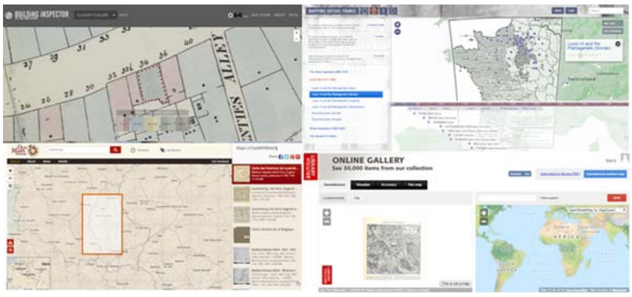


Figure 1: show screenshot of the different models of historical raster representation

Influences of raster representations

After such extensive treatment, to digitally transform the paper map into a web map, it becomes relevant to consider notions of materiality between the source and the transformed map. The "materiality" of the raster map is influenced by the transposition from paper to digital, with the associated loss of integrity. Take the action of cropping the map, observed in two thirds of all the rasters. The deletion of contextual map elements implies a certain loss of meaning from the source. Lost information commonly include: title, coordinate system, data of production, map producer, original legend, handwritten notes, archive stamps and other information. Thus the cropping process ensures that only purely functional cartographic data are retained, whilst the social document context is neglected.

A new type of digital materiality is derived from the transformation of the paper map to the digital into a " *slippy map*", where the historical raster map is cut up into a sets of map tiles that provide small images which are seamlessly joined together (in the same way Google Maps API and others provide users access to contemporary basemaps). By cropping the historical map to show only the cartographic map many map sheets from the same original paper map series can be stitched together to provide a detailed view across a wider geographical extent without having overlapping boundaries or edge effects of the different map sheets that would result otherwise. Of course such a large map with street level detail would not have been possible to draw using traditional paper maps and is facilitated only by advances in technology and web-mapping interfaces. Once the georeferenced maps are tiled, the slippy map enables users to zoom and pan effortlessly. There are certain advantages: (1) users can explore an entire city, region or continent as one seamless entity; (2) viewing of the map is instantaneous as tiling methods do not require lengthy loading times (which can be the case if providing high resolution georeferenced raster maps that are not tiled); (3) users are provided with new types of interactions for engaging with the map and importantly (4) users are provided with digital transformation to enable new ways of consuming "original" historic material.

Results: Exploring point representations

Vector data, especially point objects, were by far the most common representation of historical data in Geohumanities projects. Nearly all of the samples, almost ninety percent of the projects (n=26) used points for their digital conceptualisations of historical phenomena, ideas or events, see figure 2. Moreover, it was particularly common to simultaneously use both raster and point data structures within the interface. Whereby, the historical map in raster form was adopted as the basemap, ie a reference map with the point data provided users with representations of more specific data. Such a mix was observed in 22 projects. A small number of projects also made use of the line (26%) and area (33%) geometry objects that are available as part of the vector data format. No projects used text as a way of representing information (except to label objects label objects).

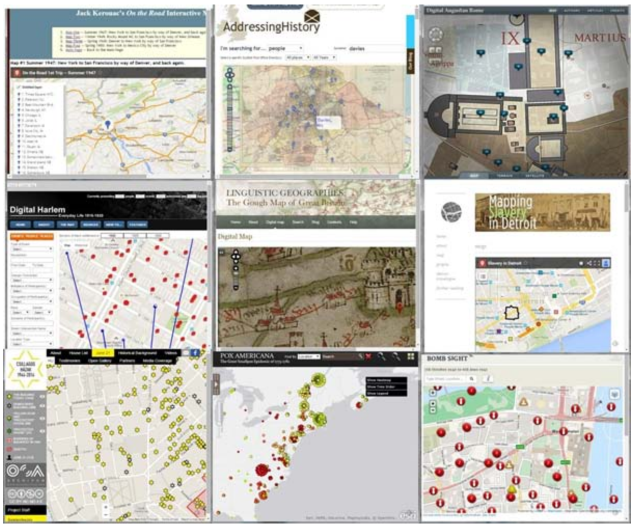


Figure 2: Example of map interfaces uses point data

Influences of Point for representing historical data

The point data object is a ubiquitous form of representation. It is a pervasive feature of geohumanities web-mapping projects and was observed in nearly all the sample projects (almost 90%). The point is used to represent a digital conceptualisation of historical phenomena, ideas, places, events etc. Despite the ubiquitous use of this data model, it is not without issues.

Given that a point marks an exact location, it was often observed that points (miss)represented phenomena that ranged in scale from a building, city and region or wider, sometimes representing all in the same data layer. Such confusing representations were exacerbated by (1) choice of representation - the use of the map pin which implies certainty in location; (2) failure to represent uncertainty in data therefore implying absolute facts – that are actually ambiguous (3) low usage of icons as map symbols reducing the cognitive understanding of the map. The point encapsulates a sense of location accuracy, precision and factual detail that often represented fuzzy historical constructs at best. Thus, in much the same way as text is transposed into digital form (Kirschenbaum, 2001; Hayles, 2005; Drucker, 2011), the digital map transformations and resulting cartographic representations influence the meaning of the mapped data and how it is read, interpreted and understood. By failing to consider carefully the cartographic choices in how the digitised historic data are embodied, it is likely that the cognitive load placed on the user and the efforts required to extract meaning from the maps are significantly increased. It also will strongly influence how these digital translations are interpreted and understood by the user.

Thoughts and Conclusions

This evaluation is a starting point for discussions between humanities scholars and geo specialists, designers, information scientists and human computer interactions experts. It is obvious that digital transformations of historical data into digital geodata world provides many benefits, not least that associated with pattern recognition and presentation of data but it is not without its limitations. Initial results indicate that existing models and cartographic representations of georeferenced historical data need to be extended. It is plausible to suggest that the point data structure is inadequate for representing complex historical phenomena or broad geographical concepts but due to the lack of alternatives it is widely accepted as standard practice. We will reflect on the point as an appropriate mechanism for representing complex historical data? Explore the extent digital representations should stay true to original historic map evidence? And investigate how *digital map translations* influence understanding and aid/hinder interpretation?

Bibliography

Drucker, J. (2011). "Diagrammatic Writing". In *Materialities of Text: Between the Codex and the Net,* edited by Sas Mays and Nicholas Thoburn. New Formations: A Journal of culture, theory and politics, **78**(1): 5-6.

Hayles, K. (2005). *My Mother Was a Computer*, University of Chicago Press. http://www.press.uchicago.edu/Misc/Chicago/321487.html .

Kirschenbaum, M. G. (2001)."Materiality and Matter and Stuff: What Electronic Texts Are Made Of". In *Electronic Book Review* (ebr), edited by Joseph Tabbi et al. http://www.electronicbookreview.com/thread/electropoetics/sited .

Authorship Attribution Using Different Languages

Patrick Juola
juola@mathcs.duq.edu
Duquesne University, United States of America

George Mikros
gmikros@isll.uoa.gr
National and Kapodistrian University of Athens, Athens, Greece

For many at this conference, stylometry and authorship attribution need little introduction; the determination of who wrote a document by looking at the writing style is an important problem that has received much research attention. Research has begun to converge on standard methods and procedures (Juola, 2015) and the results are increasingly acceptable in courts of law (Juola, 2013).

The most standard experiment looks something like this: collect a training set (aka "known documents," KD) representative of the documents to be analyzed (the testing set, aka "questioned documents," QD) and extract features from these documents such as word choice (Burrows, 1989; Binongo, 2003) or character n-grams (Stamatatos, 2013). On the basis of these features, the QD can be classified -- for example, if Hamilton uses the word "while" and Madison uses the word "whilst" (Mosteller and Wallace, 1963) a QD that doesn't use "while" is probably Madisonian.

... unless it's not in English at all, in which case, neither word is likely to appear. The need for the KD to represent the QD fairly closely is one of the major limitations on the use of this experimental methodology. By contrast, the authorial mind remains the same irrespective of the language of writing. In this paper, we report on new methods based on cross-linguistic cognitive traits that enables documents in Spanish to be attributed based on the English writings of the authors and vice versa. Specifically, using a custom corpus scraped from Twitter, we identify a number of features related to the complexity of language and expression, and a number of features related to participation to Twitter-specific social conventions.

We first identified (by manual inspection) a set of 14 user names that could be confirmed to have published tweets in both English and Spanish. Once our user list had been collected, we scraped the Twitter history of each user to collect between 90 and 1800 messages ("tweets") from each user and used the detectlanguage.com server to identify automatically the language of each tweet.

A key problem is feature identification, as most features (e.g. function words or character n-grams) are not cross-linguistic. For this work, we have identified some potentially universal features. One of the most long-standing (de Morgan, 1851) features proposed for authorship analysis is complexity of expression, as measured variously by word length, distribution of words, type/token ratio, and so forth. We used thirteen different measures of complexity that have been proposed (largely in the quantitative linguistics literature) to create a multivariate measure of complexity that persists across languages. Similarly, we identified three specific social conventions (the use of @mentions, #hashtags, and embedded hyperlinks, all measured as percentage of occurrence) that people may or may not participate in. Our working hypothesis is that people will use language in a way that they feel comfortable with, irrespective of the actual language. Hence, people who use @mentions in English will also do so in Spanish. Similarly, people who send long tweets in English also do so in Spanish, people who use big words in English also do so in Spanish, people who use a varied vocabulary in English also do so in Spanish, and of course vice versa.

We were able to show, first, that the proposed regularities do, in fact, hold across languages, as measured by cross-linguistic inter-writer correlations. (Thus, we also showed that our working hypothesis is confirmed, at least for these traits). Second, we showed via cluster analysis that these measures are partially independent from each other, and thus they afford a basis for a stylistic vector space. (Juola and Mikros, under review). This potentially enables ordinary classification methods to apply. The results reported here show that, in fact, they do.

To do this, we apply normal classification technology (support vector machines using a polynomial kernel) to the vector space thus constructed. We first broke each individual collection into 200 word sections (thus conjoining multiple tweets). Each section was measured using each complexity feature and then raw values were normalized using z-scores [thus a completely average score would be zero, while a score at the 97th percentile would be approximately 2.0; this is similar to Burrows' Delta (Burrows, 1989)]. For our first experiment, the English sections were used to create a stylometric vector space, then the Spanish sections were (individually) embedded in this space and classified via SVMs. For our second experiment, the languages were reversed, classifying English sections based on Spanish stylometric space. Since SVM with polynomial kernel is a three parameter model, we optimized the classifier's performance using a grid-search parameter tuning and comparing 3 different values for each of the three parameters (totaling 3^3 models). The classifier's performance was evaluated using a 10-fold cross-validation scheme and the best single language model was used for predicting the authorship of the texts written in the other language from the same authors.

This resulted in 2652 attempts to predict authorship of individual 200 word sections in Spanish, and another 1922 attempts in English, classified across fourteen potential authors. Baseline (chance) accuracy is therefore 1/14 or 0.0714 [7.14%].

Using the English data to establish the stylometric space and the Spanish samples to be attributed yielded an accuracy of 0.095, a result above baseline but not significantly so. By contrast, embedding English data into a Spanish space yielded an accuracy of 0.1603, more than double the baseline. This result clearly establishes the feasibility of cross-linguistic authorship attribution, at least at the proof of concept level. Experiments are continuing, both to establish clearer statistical results, and also to evaluate the additional effectiveness of the Twitter-specific social conventions as features.

We believe this result to be the first recorded instance of using training data from one language to attribute test data from another language using a formal, statistical attribution procedure. This is a very difficult dataset using an extremely small set of predictive variables, and the samples (200 words) are very small (Eder, 2013). In light of these issues, the relatively low (in absolute terms) accuracy may still represent a major step forward.

Like many research projects, these results pose as many questions as they answer. Why is English->Spanish easier than Spanish->English? What other types of language-independent feature sets could be developed, and how would performance compare? Do these results generalize to different language pairs, or to different genres than social media and Twitter in particular? What additional work will be necessary to turn this into a practical and useful tool? Can this generalize to other authorial analysis applications such as profiling (of personality or other attributes)?

Further research will obviously be required to address these and other issues. In particular, this study is obviously only a preliminary study. More language pairs are necessary (but finding active bilinguals on Twitter is difficult). Studies of other genres than tweets would be informative, but again corpus collection is problematic. We acknowledge that the current accuracy is not high enough to be useful. For the present, however, the simple fact that cross-linguistic authorship attribution can be done and has been done, remains an important new development in the digital humanities.

Bibliography

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, **16**(2): 9–17.

Burrows, J. F. (1989).`An ocean where each kind...': Statistical analysis and some majordeterminants of literary style. *Computers and the Humanities*, **23**(4-5): 309-21.

Eder, M. (2013). Does size matter? Authorship attribution, short samples, big problem. *Digital Scholarship in the Humanities*, **30**(2): 167–82.

De Morgan, A. (1851). Letter to Rev. Heald 18/08/ 1851. In Elizabeth, S. and Morgan, D. (eds), *Memoirs of Augustus de Morgan by His Wife Sophia Elizabeth de Morgan with Selections from His Letters*. Cambridge: Cambridge University Press.

Juola, P. (2013). Stylometry and immigration: A case study. *Journal of Law and Policy*, **21**(2): 287–98.

Juola, P. (2015). The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*.

Juola, P. and Mikros, G. (under review). Cross-Linguistic Stylometric Features: A Preliminary Investigation. Ms. Submitted to *JADT 2016*.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist. Reading*, MA: Addison-Wesley.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, **21**(2): 420-40.

Regional Classification of Traditional Japanese Folk Songs from the Chugoku District

Akihiro Kawase
kawase.ak@gmail.com
National Institute for Japanese Language and Linguistics, Japan

Introduction

This study aims to grasp the regional differences in the musical characteristics inherent in the traditional Japanese folk songs of the Chugoku district (the westernmost region of Japan's largest island of Honshu) by extracting and comparing the characteristics of each area by conducting quantitative analysis in order to promote digital humanities research on traditional Japanese folk songs.

We have sampled and digitized the five largest song genres within the music corpora of the *Nihon Min'yo Taikan* consisting of 1,794 song pieces from 45 Japanese prefectures, and have clarified the following three points by extracting and comparing their respective musical patterns (Kawase and Tokosumi 2011): the most important characteristics in the melody of Japanese folk songs is the transition pattern, which is based on an interval of perfect fourth pitch; regionally adjacent areas tend to have similar musical characteristics; and the differences in the musical characteristics almost match the East-West division in the geolinguistics or in the folkloristics from a broader perspective.

However, to conduct more detailed analysis in order to empirically clarify the structures by which music has spread and changed in traditional settlements, it is necessary to expand the data and do comparisons based on the old Japanese provinces (ancient administrative units that were used under the ritsuryo system before the modern prefecture system was established).

Procedure

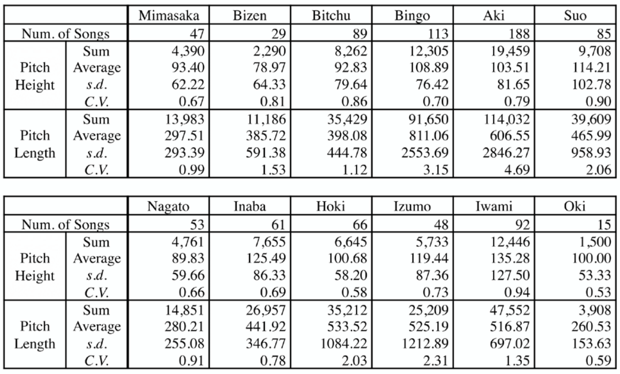
In order to digitize the Japanese folk song pieces, we generate a sequence of notes by converting the music score into MusicXML file format. We devised a method of digitizing each note in terms of its relative pitch by subtracting the next pitch height for a given MusicXML. It is possible to generate a sequence T that carries information about the pitch to the next note: $T = (t_1, t_2, \dots, ti, \dots, tn)$. An example of the corresponding pitch intervals for ti can be written as shown in Table 1. We treat sequence T as a categorical time series, and execute n-gram analysis by conducting unigram, bigram, and trigram patterns to clarify major transitions and their trends in the Chugoku district.

t_i	Pitch Intervals	t_i	Pitch Intervals
0	perfect unison	7	perfect fifth
1	minor second	8	minor sixth
2	major second	9	major sixth
3	minor third	10	minor seventh
4	major third	11	major seventh
5	perfect fourth	12	perfect octave
6	aug.fourth/dim.fifth	13	minor ninth

Table 1: Corresponding Pitch Intervals

Overview of Data

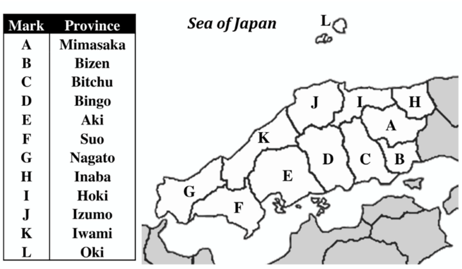
In order to quantitatively extract pitch transition patterns from Japanese folk songs, we sampled and digitized folk songs included in the *Nihon Min'yo Taikan*. We sampled all the songs in the music corpora included in the Chugoku Region volume (1966). Table 2 shows the statistics about the song pieces from each area in the Chugoku district. Figure 1 is a map of the provinces in the Chugoku district.



		Mimasaka	Bizen	Bitchu	Bingo	Aki	Suo
Num. of Songs		47	29	89	113	188	85
Pitch Height	Sum	4,390	2,290	8,262	12,305	19,459	9,708
	Average	93.40	78.97	92.83	108.89	103.51	114.21
	s.d.	62.22	64.33	79.64	76.42	81.65	102.78
	C.V.	0.67	0.81	0.86	0.70	0.79	0.90
Pitch Length	Sum	13,983	11,186	35,429	91,650	114,032	39,609
	Average	297.51	385.72	398.08	811.06	606.55	465.99
	s.d.	293.39	591.38	444.78	2553.69	2846.27	958.93
	C.V.	0.99	1.53	1.12	3.15	4.69	2.06

		Nagato	Inaba	Hoki	Izumo	Iwami	Oki
Num. of Songs		53	61	66	48	92	15
Pitch Height	Sum	4,761	7,655	6,645	5,733	12,446	1,500
	Average	89.83	125.49	100.68	119.44	135.28	100.00
	s.d.	59.66	86.33	58.20	87.36	127.50	53.33
	C.V.	0.66	0.69	0.58	0.73	0.94	0.53
Pitch Length	Sum	14,851	26,957	35,212	25,209	47,552	3,908
	Average	280.21	441.92	533.52	525.19	516.87	260.53
	s.d.	255.08	346.77	1084.22	1212.89	697.02	153.63
	C.V.	0.91	0.78	2.03	2.31	1.35	0.59

Table 2: Basic Statistics for Number of Songs for Each Province



Mark	Province
A	Mimasaka
B	Bizen
C	Bitchu
D	Bingo
E	Aki
F	Suo
G	Nagato
H	Inaba
I	Hoki
J	Izumo
K	Iwami
L	Oki

Figure 1: Geographical Divisions of the Chugokuk District Under the Old Province System

Results

Frequency of the First Transition

Figure 2(a) shows the usage frequency of the first transition patter (unigram). The graph implies that pitch transitions occur almost equally in both the ascending and descending directions. Figure 2(b) is a cumulative relative frequency diagram of the first transitions for ascending and descending order that confirms the trends for each interval. The profile shows that most of the voice range does not extend beyond the interval of perfect fifth (±7).

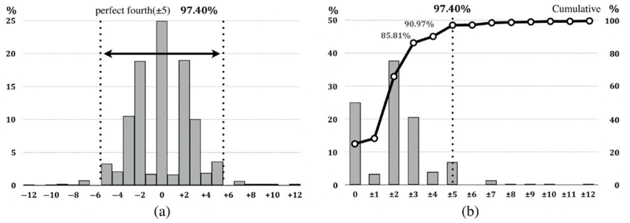


Figure 2: (a) First Transition Frequency and (b) Cumulative Relative Frequency Diagram

Based on studies of musical psychology, step-by-step changes in pitch height (in other words, pitch interval transitions that always involve ascending or descending order) are an indispensable factor by which humans perceive melodic patterns (Scruton 1997). Therefore, to extract the structures buried under repetition of notes, we will exclude the 0 intervals from sequence T in order to seek the pure changes of ups and downs within the melody.

Overall Summary of Bigram Patterns

Table 3(a) shows the top 40 bigram patterns. These components do not exceed the range of the perfect fifths (±7), and the values of the chains (intervals) also add up to a pitch interval between -5 and +5. In addition, we find that transition patterns that add up to 0 ($t_i + t_{i+1} = 0$), transition patterns that add up to ±5 ($t_i + t_{i+1} = \pm 5$, which is the interval of the perfect fourth), and patterns that include the interval of the perfect fourth in either component (e.g. patterns that form (t_i, t_{i+1}) = (±5, *), (*, ±5)) also appear with high frequency.

(a)

Rank	t_i	t_{i+1}	Sum	Num.	Rank	t_i	t_{i+1}	Sum	Num.
1	+2	-2	0	6,642	21	-3	+5	+2	712
2	-2	+2	0	6,222	22	-4	+2	-2	672
3	-2	-3	-5	5,089	23	-2	-5	-7	664
4	+3	+2	+5	4,773	24	-5	-2	-7	662
5	-3	-2	-5	4,189	25	+4	-2	+2	584
6	+2	+3	+5	3,968	26	-2	+4	+2	569
7	-3	+3	0	3,695	27	-2	+5	+3	553
8	-2	-2	-4	3,500	28	-1	+1	0	492
9	+2	+2	+4	3,135	29	-4	-1	-5	486
10	+3	-3	0	3,074	30	-1	-2	-3	461
11	+2	+5	+7	1,130	31	+2	+1	+3	447
12	+5	+2	+7	988	32	+1	+4	+5	403
13	+5	-2	+3	932	33	+4	-4	0	372
14	+2	-5	-3	879	34	-4	+4	0	331
15	-5	+3	-2	834	35	-5	+5	0	307
16	-5	+2	-3	766	36	-7	+2	-5	291
17	+2	-4	-2	766	37	-4	-3	-7	289
18	+3	-5	-2	753	38	-1	-4	-5	256
19	+5	-3	+2	731	39	+5	-5	0	248
20	+1	-1	0	714	40	+4	+1	+5	245

(b)

Rank	t_i	t_{i+1}	t_{i+2}	Sum	Num.	Rank	t_i	t_{i+1}	t_{i+2}	Sum	Num.
1	-2	-3	+3	-2	2,255	21	+2	+2	+3	+7	1,106
2	-2	+2	-2	-2	2,151	22	+3	+2	+3	+8	887
3	+2	-2	-3	-3	1,992	23	+3	-3	+3	+3	863
4	+2	+3	+2	+7	1,970	24	-3	-2	-3	-8	795
5	-3	+3	+2	+2	1,950	25	+5	+2	-2	+5	473
6	+2	-2	+2	+2	1,893	26	-2	-3	+5	0	466
7	-2	-3	-2	-7	1,784	27	+2	-5	+3	0	460
8	+3	+2	-2	+3	1,758	28	-2	+2	+5	+5	419
9	+2	-2	-2	-2	1,702	29	-5	+3	+2	0	414
10	+3	-3	-2	-2	1,666	30	+2	-4	+2	0	408
11	-3	-2	+2	-3	1,536	31	+5	-2	-3	0	399
12	-2	-2	+2	-2	1,526	32	-2	+2	-5	-5	397
13	+2	+2	-2	+2	1,459	33	-3	+5	-2	0	397
14	-2	-2	-3	-7	1,425	34	+2	+5	-2	+5	379
15	+2	+3	-3	+2	1,387	35	+5	-3	-2	0	378
16	-2	+2	+2	+2	1,304	36	-2	+4	-2	0	372
17	+3	+2	+2	+7	1,213	37	+3	+2	+5	+10	366
18	-3	-2	-2	-7	1,177	38	+5	-2	+2	+5	365
19	-2	+2	+3	+3	1,159	39	+4	-2	-2	+4	364
20	-3	+3	-3	-3	1,140	40	+2	+2	-4	0	359

Table 3: Frequency of Occurrence of (a) Bigrams and (b) Trigram (Top 40 Patterns)

Overall Summary of Trigram Patterns

Table 3(b) shows the top 40 trigram patterns. These components do not exceed the range of perfect fourths (±5), and the feature that stands out is that some of the top-ranked patterns are patterns with a single pitch transition attached to bigram. Most of the transition patterns fall in the range of the intervals of perfect fifths (±7). Furthermore, there are many high-ranked patterns in which the beginning two transitions or the last two transitions add up to form either the same degree of pitch ($t_i + t_{i+1} = 0$ or $t_{i+1} + t_{i+2} = 0$) or a perfect fourth pitch ($t_i + t_{i+1} = \pm 5$ or $t_{i+1} + t_{i+2} = \pm 5$).

Discussion

Koizumi's Tetrachord Theory

To speed things along, we will mention Koizumi's tetrachord theory (1958). The tetrachord is a unit consisting of two stable outlining tones (nuclear tones) with the interval of a perfect fourth pitch, and one unstable intermediate tone located between them. Depending on the position of the intermediate tone, four different types of tetrachords can be formed (Table 4).

Type	Name	Pitch Interval
I	*Min'yo*	minor third (3) + major second (2)
II	*Miyako bushi*	minor second (1) + major third (4)
III	*Ritsu*	major second (2) + minor third (3)
IV	*Ryukyu*	major third (4) + minor second (1)

Table 4: For Basic Types of Tetrachords

Using a bigram model representing pitch transitions, all four types of tetrachords can be expressed as follows in ascending order: *min'yo* (+3, +2), *miyako bushi* (+1, +4), *ritsu* (+2, +3), and *ryukyu* (+4, +1). Depending on the positions of the three initial pitches in a tetrachord, six transition patterns can be considered in perceiving a tetrachord in two steps (bigram). Therefore, the amount of tetrachords within two steps can be obtained by counting the pairs of 24 transition patterns in sequence T.

	Min'yo tetrachord						*Miyako bushi* tetrachord					
	(+3, +2)	(+5, -2)	(+2, -5)	(-3, +5)	(-2, -3)	(-5, +3)	(+1, +4)	(+5, -4)	(+4, -5)	(-1, +5)	(-4, -1)	(-5, +1)
Mimasaka	266	16	49	12	211	62	16	1	3	3	6	3
Bizen	124	23	17	21	126	18	9	6	3	6	24	0
Bitchu	435	87	50	72	471	46	39	2	7	12	52	11
Bingo	665	108	135	72	641	128	50	5	6	5	40	4
Aki	981	179	173	149	1,030	151	85	13	19	22	95	20
Suo	416	83	58	70	485	78	60	15	23	24	84	26
Nagato	204	49	36	39	210	39	60	3	5	10	64	7
Inaba	335	54	67	35	362	63	35	10	7	17	63	5
Hoki	382	77	57	34	440	66	26	1	2	5	23	0
Izumo	352	63	73	61	371	70	6	2	0	2	9	0
Iwami	552	169	151	134	661	106	12	3	0	13	21	4
Oki	61	24	13	13	81	7	5	0	1	0	5	1

	Ritsu tetrachord						*Ryukyu* tetrachord					
	(+2, +3)	(+5, -3)	(+3, -5)	(-2, +5)	(-3, -2)	(-5, +2)	(+4, +1)	(+5, -1)	(+1, -5)	(-4, +5)	(-1, -4)	(-5, +4)
Mimasaka	191	23	43	13	167	42	15	7	11	0	16	3
Bizen	99	13	20	12	84	20	5	1	0	0	11	0
Bitchu	294	50	48	35	316	47	27	3	2	5	29	0
Bingo	572	90	100	72	558	126	24	2	1	4	15	3
Aki	841	156	170	108	830	196	49	7	17	10	32	13
Suo	320	114	62	83	411	52	53	5	15	11	61	11
Nagato	170	36	43	20	165	28	15	1	3	1	26	3
Inaba	337	47	43	57	377	43	27	1	5	2	28	3
Hoki	296	48	46	20	355	41	18	1	0	1	20	0
Izumo	274	44	51	31	273	37	1	2	0	1	4	1
Iwami	499	99	112	93	576	122	11	3	3	6	14	1
Oki	75	11	15	9	77	12	0	0	0	0	0	0

Table 5: Number of Transition Patterns for the Four Tetrachords

Province Similarities Based on the Frequency of Occurrence of Tetrachords

Table 5 shows the frequency that the 24 transition patterns for the four tetrachords appear for all 12 provinces. Here, we did hierarchical cluster analysis on the 24 transition patterns for each province to objectively demonstrate similarities in the 12 provinces. The dendrogram in Figure 3 shows hierarchical cluster analysis results. When calculating distances between each element, we normalized the frequency that the tetrachords appear, and used the Euclidean distance and the algorithm from the Ward method.

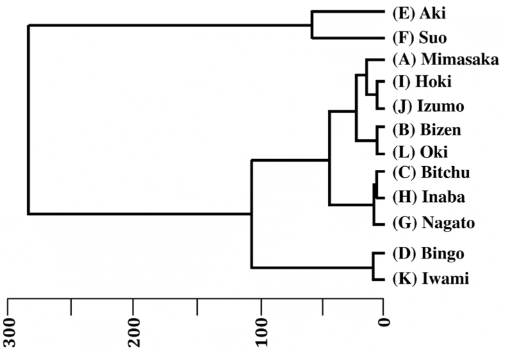


Figure 3: Cluster Analysis Results

If we look for $h = 60$ (a height where there are three vertical lines) and trace them to the individuals, this partition of three clusters separates the 12 provinces into C_1 = {Aki, Suo}, C_2 = {Mimasaka, Bizen, Bitchu, Nagato, Inaba, Hoki, Izumo, Oki}, and C_3 = {Iwami, Bingo}, which almost classifies the Chugoku district into geographically close groups. Meanwhile, it also stands out that only Nagato is separately geographically isolated.

Sea Route Connecting the Provinces

The cluster analysis results did not turn out to match the famous two ancient divisions based on the main road: San'yo = {Mimasaka, Bizen, Bitchu, Bingo, Aki, Suo, Nagato} and San'in = {Inaba, Hoki, Izumo, Iwami, Oki}. However, instead it is possible to consider each cluster based on sea routes (Figure 4).

Since ancient times, the Chugoku district has been an area travelers passed through when traveling from the Kansai district (a center of regional culture in Japan) to get to Kyushu region, which had close relations with China and Korea. Since there wasn't much flat land and there are many mountain ranges, sea routes played an important role in transportation.

The existence of the Ota River, a major river and also a sociocultural and economical foundation in the Chugoku district, supports the grouping of the two provinces in C_1. From an oceanic commerce point of view, the north and south sea routes that link the Nagato and Kansai regions also support the grouping of providences in C_2. It is also possible to say that the Gonokawa River, which functioned as a traffic relay point between San'yo and San'in, supports the grouping of provinces in C_3. In this way, we find that rivers have important functions that exert musical influence as well as indicate boundaries between districts and geographical areas.

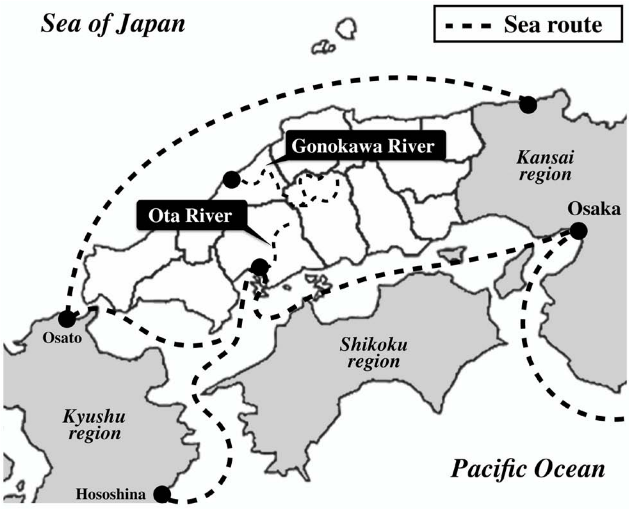


Figure 4: Sea Routes Map of the Chugoku District

Conclusions

In this paper, we digitized the melodies of traditional Japanese folk songs in the Chugoku district, and quantitatively analyzed its pitch transition patterns using n-gram modeling, and did a classification experiment based on the frequency of occurrence of the tetrachords to see the differences in each province. As a result, we constructed the possibility that the melodic features in the Chugoku district spread by land and sea routes based on actual music data analysis. However, we should be able to describe the relationships influencing musical culture between regions in detail if we develop this analysis on a nationwide scale. In further research, it is possible to clarify the structural commonalities and differences between areas by conducting analysis on musical corpora nationwide including folk song pieces from neighboring regions such as the Kyushu, Shikoku, and Kansai regions, which has not really been pursued in existing humanities research fields. We believe this will empirically clarify the musical culture phenomena by which music spreads and changes.

Bibliography

Kawase, A. and Tokosumi, A. (2011). Regional classification traditional Japanese folk songs, *International Journal of Affective Engineering*, **10**(1): 19-27.

Koizumi, F. (1958). *Studies on Traditional Music of Japan 1*, Ongaku no tomosha.

MusicXML. http://www.musicxml.com/for-developers/ (accessed 16 February 2016).

Nihon Hoso Kyokai (1944-1993). *Nihon Min'yo Taikan(Anthology of Japanese Folk Songs).*

Scruton, R. (1997). *The Aesthetics of Music*, Oxford University Press.

Authorship Verification with the Ruzicka Metric

Mike Kestemont
mike.kestemont@gmail.com
University of Antwerp, Belgium

Justin Stover
justin.stover@classics.ox.ac.uk
University of Oxford, United Kingdom

Moshe Koppel
moishk@gmail.com
Bar-Ilan University, Israel

Folgert Karsdorp
fbkarsdorp@fastmail.nl
Meertens Institute, The Netherlands

Walter Daelemans
walter.daelemans@uantwerpen.be
University of Antwerp, Belgium

Introduction

Authorship studies have long played a central role in stylometry, the popular subfield of DH in which the writing style of a text is studied as a function of its author's identity. While authorship studies come in many flavors, a remarkable aspect is that the field continues to be dominated by so-called 'lazy' approaches, where the authorship of an anonymous document is determined by extrapolating the authorship of a document's nearest neighbor. For this, researchers use metrics to calculate the distances between vector representations of documents in a higher-dimensional space, such as the well-known Manhattan city block distance. In this paper, we apply the minmax metric – originally proposed in the field of geobotanics – to the problem of authorship attribution and verification. Comparative evaluations across a variety of benchmark corpora show that this metric yields better, as well as more consistent results than previously used metrics. While intuitively simply, this metric generally displays a regularising effect across different hyperparametrizations, and allows the more effective use of larger vocabularies and sparser document vectors. In particular the metric seems much less sensitive than its main competitors to (the dimensionality of) the vector space model under which the metric is applied.

Most authorship studies in computer science are restricted to present-day document collections. In this paper, we illustrate the broader applicability of the minmax metric by applying it to a high-profile case study from Classical Antiquity. The 'War Commentaries' by Julius Caesar (*Corpus Caesarianum*) refers to a group of Latin prose commentaries, describing the military campaigns of the world-renowned Roman statesman Julius Caesar (100-44 BC). While Caesar must have authored a significant portion of these commentaries himself, the exact delineation of his contribution to this important corpus remains a controversial matter. Most notably, Aulus Hirtius – one of Caesar's most trusted generals – is sometimes believed to have contributed significantly to the corpus. Thus, the authenticity and authorship of the Caesarian corpus is a philological puzzle that has persisted for nineteen centuries. In our paper, we shed new light on this matter.

Benchmarking

To properly evaluate the performance of the novel Ruzicka minmax metric, we turn to a publicly available benchmark corpora: the multilingual datasets (Dutch, English, Greek, and Spanish) used by the 2014 track on authorship verification in the PAN competition on uncovering plagiarism, authorship, and social software misuse. This track focused on the "open" task of authorship verification (as opposed to the closed set-up of authorship verification). Each dataset holds a number of "problems", where given (a) at least one training text by a particular target author, (b) a set of similar mini-oeuvres by other authors, and (c) a new anonymous text, the task is to determine whether or not the anonymous text was written by the target author. A system must output for each of the problems a real-valued confidence score between 0.0 ("definitely not the same author") and 1.0 ("definitely the same author"). By outputting the value of 0.5, a system can specify that it was not able to solve a problem. For each dataset, a fully independent training and test corpus are available (i.e. the problems, nor authors and texts in both sets do not overlap). Systems are eventually evaluated using two scoring metrics which were also used at the PAN: the established AUC-score, as well as the so-called c@1, a variation of the traditional accuracy-score, which gives more credit to systems that decide to leave some difficult verification problems unanswered. In the full paper, we offer a complete evaluation of all datasets: for the sake of brevity, this paper is restricted to a representative selection of results.

As common in text classification research, we vectorize the datasets into a tabular model, under a 'bag-of-words' assumption, which is largely ignorant of the original word order in document. Unless reported otherwise, we use character tetragrams below (Koppel et al., 2014), which yield generally acceptable results across corpora. We experiment with a number of different vector space models, the results of which can be summarized as follows:

- plain *tf* (where simple relative frequencies are used);
- *tf-std*, where the *tf*-model is scaled using a feature's standard deviation in the corpus (cf. Burrows's Delta: Burrows, 2002);
- *tf-idf*, where the *tf*-model is scaled using a feature's inverse document-frequency (to increase the weight of rare terms).
- …

In our experiments, we focus on the Ruzicka 'minmax' distance metric, a still fairly novel algorithm in the field of stylometry. Just as the Euclidean or Manhattan distance, this metric will calculate a real-valued distance score between two document vector A and B as follows:

$$minmax(\vec{A}, \vec{B}) = 1 - \left(\frac{\sum_{i=1}^{n} \min(\text{tf}(A_i), \text{tf}(B_i))}{\sum_{i=1}^{n} \max(\text{tf}(A_i), \text{tf}(B_i))} \right)$$

While the formula below uses the tf-model, the Ruzicka distance can of course be easily applied to other vector space models too. In our paper, we will offer a intuitive assessment of the desirable properties of this metric (e.g. in comparison to Burrows's Delta).

General Imposters Framework (GI)

In our experiments, we make amongst others use of the General Imposters Method, a bootstrapped approach to authorship verification which has recently yielded excellent results. Fitting the verifier on the train data involves two steps. First, we calculate a distance score for the anonymous document in each problem, using Algorithm 1, in order to determine whether the anonymous text was written by the target author specified in the problem:

```
Algorithm 1: General Imposters Method
input  : $x$, an anonymous document; $T = \{T_1, \ldots, T_n\}$, a
         set of document by the target author;
         $I = \{I_1, \ldots, I_n\}$, a set of document by other
         authors;
output: $0 <= score <= 1$
Set $score = 0.0$;
for $i \leftarrow 1$ to $k$ do
    Randomly select $rate$% of the available features;
    Randomly select $m$ imposters from $I$ as $I'$;
    if $min(dist(T, x)) < min(dist(I', x))$ then
        $score = score + 1/k$ ;
end
Return $1 - score$;
```

Thus, during *k* iterations (default 100), we randomly select a sample (e.g. 50%) of all the available features in the data set. Likewise, we randomly select *m* 'imposter' documents (default 30), which were not written by the target author. Next, we use a *dist()* function to assess whether the anonymous text is closer to any text by the target author than to any text written by the imposters. Here, *dist()* represents a regular, geometric distance metric, such as the Manhattan or Ruzicka metric. The score returned by Algorithm 1 has been characterized as a 'second-order' metric, because it does not rely on the rather comparison of document vectors. The general intuition here, is that we do not just calculate how different two documents are; rather we test whether the stylistic differences between them are consistent (a) across many different feature sets, and (b) in comparison to other randomly, sampled documents.

In the second stage, we attempt to optimize the distance scores returned by Algorithm 1, in the light of the specific evaluation measures used. We apply a score shifter (Algorithm 2), which attempts to define a 'grey zone' where the results seem too unreliable to output a score (cf. c@1):

```
Algorithm 2: Score shifting
input  : $scores$; $0 <= p_1 <= 1$;
$0 <= p_2 <= 1$;
output: $shifted\_scores$
Set $shifted\_scores = list()$;
for $score$ in $scores$ do
    if $score < p_1$ then
        $new\_score = rescale(score)$ to $range(0.0, p_1)$;
    else if $p_1 <= score <= p_2$ then
        $new\_score = 0.5$;
    else
        $new\_score = rescale(score)$ to $range(p_2, 1.0)$;
    end
    Append $new\_score$ to $shifted\_scores$;
end
Return $shifted\_scores$;
```

Through a grid search of different values between 0 and 1 for p1 and p2, we determine the settings which yield the optimal AUC x c@1 on the train data. In Fig. 1, we plot the optimal results which could be obtained on the train problems in the data set of Dutch Essays, for a specific combination of a metric and a vector space model. We ran the experiment 20 times, with increasing vocabulary truncations (e.g. the 1000 most frequent tetragrams). The results demonstrate how the Ruzicka minmax metric returns the most stable results across the experiments and clearly has a regularizing effect across different hyperparametrizations. In the full paper, we will present a complete evaluation of this system on all the PAN datasets, which in most cases yields surprisingly competitive scores on the test data, even without much corpus-specific parameter tuning. In the table below, we show the test results for Dutch essays corpus in terms of the AUC x c@1. The best combination reaches a AUC x c@1 of 0.886 on the test data (combination of *minmax* and *std*), whereas the best individual system submitted to PAN 2014 only reached 0.823 on that test dataset. Using randomized significance tests, we will additionally demonstrate the regularizing effect of the Ruzicka distance across vector spaces; its strong performance is also evident from Table 1.

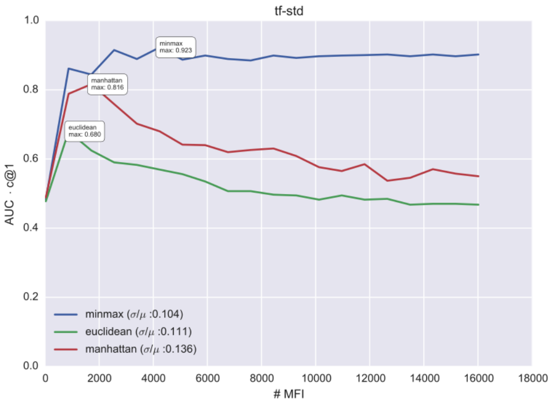


Figure 1: Optimal results on train corpus

Vector Space / Metric	Euclidean	Manhattan	Minmax
Tf	0.676	0.698	0.837
Tf-Idf	0.720	0.750	0.854
Tf-Std	0.614	0.701	0.886

Table 1: Final test results (AUC x C@1)

Corpus Caesarianum

To further illustrate the applicability of the Ruzicka metric for authorship problems in traditional philology, we also report a stylometric case study concerning the *Corpus Caesarianum*. This *Corpus* is a group of five commentaries Caesar's military campaigns:

- *Bellum Gallicum*, the conquest of Gaul, 58 to 50 BC;
- *Bellum civile*, the civil war with Pompey, 49 to 48 BC;
- *Bellum Alexandrinum*, the campaigns in Egypt etc., 48 to 47 BC;
- *Bellum Africum*, the war in North Africa, 47 to 46 BC
- *Bellum Hispaniense*, a rebellion in Spain, 46 to 45 BC

The first two commentaries are mainly by Caesar himself, the only exception being the final part of the *Gallic War* (Book 8), which is by Caesar's general Aulus Hirtius. Suetonius, writing a century and a half later, suggests that either Hirtius or another general, named Oppius, authored the remaining works. We will report experiments which broadly supports the Hirtius's own claim that he himself compiled and edited the corpus of the non-Caesarian commentaries. Figure 2, for instance, shows a heatmap-like visualisation, in which Hirtius's Book 8 of the *Gallic War* clearly clusters with the bulk of the *Alexandrian War* (labeled *x*).

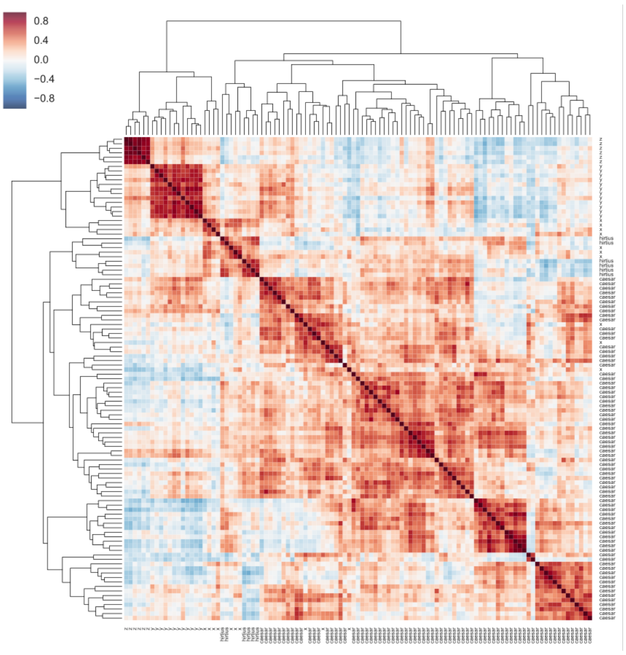


Figure 2: Minmax-based clustermap of 1000-word samples of the *Corpus Caesarianum*.

Bibliography

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**: 131-47.

Burrows. J. F. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**: 267-87.

Gaertner, J. and Hausburg, B. (2013). *Caesar and the Bellum Alexandrinum: An Analysis of Style, Narrative Technique, and the Reception of Greek Historiography*. Vandenhoeck & Ruprecht, Göttingen.

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, **65**: 178–87.

Mayer, M. (2011). Caesar and the corpus caesarianum. In Marasco, G. (ed), *Political auto-biographies and memoirs in antiquity: A Brill companion*. Brill, Leiden, pp. 189-232.

Stamatatos, E. et al. (2014). Overview of the author identification task at PAN 2014. In *Working Notes for CLEF 2014 Conference*, pp. 877-97.

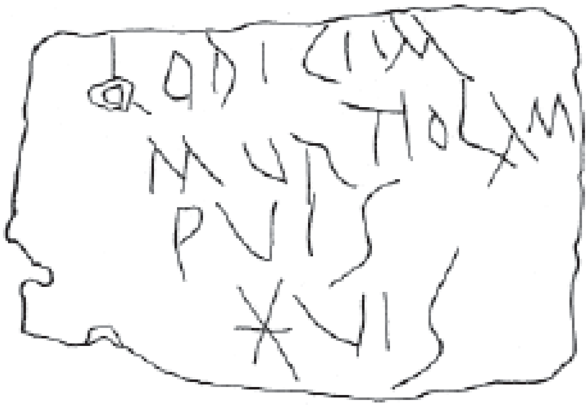
Stover, J., Winter, Y., Koppel, M. and Kestemont, M. (2016). Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the American Society for Information Science and Technology*, **66**: 239-42.

GLAMorous! Edizione Digitale Di Beni Culturali Con Contenuto Testuale, Multidisciplinarietà Ed Epigrafia Digitale.

Marion Lamé
mlame@mmsh.univ-aix.fr
Centre Camille Jullian, MMSH, CNRS, France.; Laboratorio di Cultura Digitale, Università di Pisa; Istituto di Linguistica Computazionale "A. Zampolli", CNR, Italy

Il progetto *Tesserarum Sisciae Sylloge* (TSS) mira alla digitalizzazione di più di 1100 etichette commerciali di piombo di epoca romana (I-III secolo d.C.) conservate nel Museo Archeologico di Zagabria. Hanno le dimensioni approssimattive dei nostri francobolli e portano incisioni con informazioni relative alla tintura della lana (prezzi, prodotti, ecc.).

L'edizione scientifica è a cura di Radman-Livaja, 2014 e il disegno seguente è il fronte dell'etichetta 12563 pubblicata dall'editore



Queste etichette, beni archeologici sotto la responsabilità dei GLAM (Galleries, Libraries, Archives and Museums), sono anche oggetto di studio scientifico, e quindi da considerare sotto vari punti di vista: beni culturali, fonti primarie di informazione per gli studi storici, epigrafi, *Text Bearing Object* (TBO) etc. Sia nell'interesse del pubblico che della ricerca, abbiamo realizzato, tramite l'edizione digitale di queste etichette romane, un open access a questa documentazione (Terras, 2015), facilitato da un *framework* digitale che copre gli aspetti di visualizzazione, divulgazione e informazione scientifica.

Lavori precedenti (vedi ad es. Terras, 2006 con le tavolette di Vindolanda) avevano approfondito le riflessioni riguardo alla rapresentazione digitale e alla decifrazione di un TBO simile per certi aspetti (scritura manoscritta) e sono serviti come punto di partenza in fase preparatoria. Il progetto TSS è stato, durante l'anno 2015, l'ambiente di sviluppo di un aspetto metodologico specifico dell'epigrafia digitale, che mette al centro della procedura di digitalizzazione l'oggetto archeologico, cioè la fonte primaria di informazione dello storico. Si basa sull'analisi dispositiva (*dispositive analysis*) ed è approfondito in Lamé 2015a. L'analisi dispositiva prende in considerazione i vari elementi eterogenei che partecipano al messaggio epigrafico, tenendo conto delle dimensioni sociali e culturali del TBO in quanto oggetto di studio della storia digitale. Propone un insieme di tappe e di procedure che possono includere il *crowdsourcing* (vedi Ridge 2014).

Raggiunti i primi obiettivi prefissati, utilizzando le tecnologie di *Reflectance Transforming Imaging* (RTI), siamo entrati in una fase che potremmo chiamare di *document editing* e che riguarda le metodologie del Cultural Heritage, i *Digital Autoptic Processes* (DAP), la materialità dell'oggetto, e la sua accessibilità a tutti i pubblici dell'Archeological Museum of Zagreb in modo da svolgere sia una missione scientifica che di mediazione museografica verso il suo pubblico. Tuttavia, come sottolinea Pierazzo 2015 (specialmente pp. 70-83), *document editing* (edizione del TBO) e *textual editing* (edizione del testo) sono metodi editoriali complementari: se si possono distinguere non si possono separare. Il *textual editing* è uno dei campi di ricerca più sviluppato nell'ambito delle Digital Humanities ed è ben supportato dal punto di vista tecnico, in particolare grazie allo standard TEI. Il testual editing ha raggiunto la matturità necessaria per essere abbinato e combinato agli aspetti che riguardano il documento.

Il contesto dell'epigrafia digitale del progetto TSS, che presenteremo nella sua versione 1.0, con le conclusioni metodologiche attinenti all'informatica umanistica, in occasione del convegno DH16, fornisce documenti epigrafici la cui funzione è per natura diversa dal manoscritto. Questo genera alcune conseguenze, come, ad esempio, l'assenza del concetto di pagina ma l'uso di quello meno controllabile tecnicamente di 'specchio epigrafico', ovvero la superficie inscritta; essa non viene sempre definita da un rettangolo piano, ma può assumere numerose forme geometriche, come un cipo migliaro, avere confini meno delimitati nello spazio, come nelle iscrizioni arabe, nelle quali le scritte si miscelano a motivi floreali, o come le statue-geroglifiche egiziane. Diventa inoltre importante unire materialità archeologica (*document editing*) e interpretazione del testo (*textual editing*) in una "ricostituzione diplomatica" dell'oggetto in caso di lacuna (Lamé 2015b) tramite un tool dedicato, il MarkOut, che permette di associare calco digitale, paleografia e codifica testuale al momento della decifrazione.

MarkOut è un'interfaccia web (HTML5 e Javascript) che permette di tracciare le iscrizioni su di un'immagine e successivamente di correggere le linee tramite dei punti di controllo, consentendo di ottenere facilmente una rappresentazione fedele con uno strumento, il mouse, non proprio ideale, ma alla portata di chiunque.

Ad ogni segno viene associato un simbolo di un alfabeto predefinito, o eventualmente una variante di una lettera (Unicode code point + un indentificatore della variante). L'interfaccia infine permette di gestire più strati di scrittura e nel caso siano disponibili rappresentazioni RTI, è possibile variare l'illuminazione dell'iscrizione per facilitarne l'osservazione.

Il risultato finale è un file in formato SVG, che può essere condiviso online, processato automaticamente per estrarre la trascrizione in altri formati (es. TEI), inserito in un database ad hoc, o impiegato in algoritmi che analizzino la forma delle varie lettere e la loro posizione topologica sul TBO.

Questa interfaccia è concepita per servire i vari tipi di usi, con alcune varianti nel suo design: la discussione scientifica tra esperti (MarkOut Expert), la formazione degli studenti universitari alla decifrazione di iscrizioni (MarkOut Assisted, see Lamé 2016), la mediazione museografica verso il pubblico generale del museo, adulti e bambini. Di seguito, immagine dell'etichetta 12563 con il MarkOut Assisted.

Specialmente, il MarkOut Kids, in corso di sperimentazione presso i bambini da 4 a 10 anni, permette di mettere in pratica esercizi di scrittura del programma scolastico combinato ad un contatto con l'oggetto archeologico. Il gesto continuo che consiste nel designare la forma di una lettera dell'alfabeto viene associato a quello più astratto di identificazione e riconoscimento di un segno alfabetico con la sua posizione su una tastiera e il suo inserimento in modo tale che compaia associato al disegno sullo schermo.

Il Tesserarum Sisciae Sylloge è finanziato dal Ministero della Cultura Croato, dal Comune di Zagabria e dall'European Association for Digital Humanities. È realizzato dal Laboratorio di Cultura Digitale dell'Università di Pisa con la partecipazione del Consiglio Nazionale delle Ricerche.

Bibliography

Borillo, M. (1984). *Informatique pour les sciences de l'homme: limites de la formalisation du raisonnement*. (Philosophie et langage). Bruxelles: P. Mardaga.

Lamé, M. (2015a). Primary Sources of Information, Digitization Processes and Dispositive Analysis. New York: ACM.

Lamé, M. (2015b). Scritture in contesti: il dispositivo epigrafico come veicolo di echi epigrafici. *Lexis*, **33**: 9–17.

Lamé, M., Ponchio, F., Robertson, B. and Radman-Livaja, I. (2016). Teaching (Digital) Epigraphy. Rome: La Sapienza.

Masséglia, J. (2015). The Ashmolean Latin Inscriptions Project: Bringing epigraphic research to museum visitors and schools. *Studi Umanistici – Antichistica Convegni Information Technologies for Epigraphy and Cultural Heritage*. Rome: La Sapienza, pp. 221–30.

Pierazzo, E. (2015). *Digital scholarly editing: theories, models and methods*. Farnham; Burlington: Ashgate.

Radman-Livaja, I. (2014). *Tesere Iz Siska/Plombs de Siscia*. Vol. IX/1-IX/2. (Musei Archaeologici Zagrabiensis Catalogi et Monographiae). Zagreb: Archaeological Museum in Zagreb.

Ridge, M. (2014). *Crowdsourcing Our Cultural Heritage*. Farnham: Ashgate.

Rosselli del Turco, R. (2015). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative*(8).

Terras, M. (2006). *Image to Interpretation: An Intelligent System to Aid Historians in Reading the Vindolanda Texts*. Oxford: Oxford University Press.

Terras, M. (2015). Opening Access to Collections: the Making and Using of Open Digitised Cultural Content. *Online Information Review*(Special Issue on Open Access: Redrawing the Landscape of Scholarly Communication).

Du Texte Antique À La Publication Scientifique : Outils D'Analyse Numériques Des Contenus Et Ponts Conceptuels.

Marion Lamé
mlame@mmsh.univ-aix.fr
Centre Camille Jullian, MMSH, CNRS, France.; Laboratorio di Cultura Digitale, Università di Pisa; Istituto di Linguistica Computazionale "A. Zampolli", CNR, Italy

Blandine Nouvel
nouvel@mmsh.univ-aix.fr
Centre Camille Jullian, MMSH, CNRS, France.

Relier deux outils d'indexation pour faire émerger des ponts conceptuels

Ce travail s'applique aux sciences de l'Antiquité et à l'organisation des connaissances (Gnoli et al., 2006; Gnoli, 2012; Gnoli, 2014). Son objectif est de travailler sur la mise en lien entre les annotations appliquées à des sources primaires textuelles et des métadonnées employées pour décrire la littérature scientifique qui en découle. Plus spécifiquement, il s'agit d'étudier la façon dont sont organisés, par le biais de deux instruments terminologiques, des thèmes communs pour la description de contenus relatifs à l'Antiquité. L'un est destiné à l'analyse de textes antiques (*Index* de *Memorata Poetis*) et le second à l'indexation documentaire de publications scientifiques contemporaines (Thésaurus PACTOLS). L'objectif est de créer des relations conceptuelles entre les deux instruments, bien qu'ils ne servent pas à décrire le même type d'objets initialement, pour en améliorer l'utilisation par les usagers dans le cadre de la recherche scientifique, mais aussi d'une mise à disposition à un plus large public (Insertion de liens de références dans Wikipedia). Il fait suite au travail initié par Lamé et Caputo, 2015 sur les dispositifs épigraphiques et l'emploi de cet index thématique de *Memorata Poetis*http://www.memoratapoetis.it.

Pour cet exercice de mise en relation, nous avons choisi de privilégier une thématique familière aux spécialistes de l'Antiquité et au grand public : les dieux et héros de la mythologie grecque et latine. Leurs noms sont fréquemment employés dans la description des objets archéologiques décorés, notamment la céramique ou la sculpture antique et dans l'analyse des textes anciens. Cette thématique est traitée dans les deux outils en question de façons différentes. Le sujet s'attache donc à comparer la façon dont les deux instruments présentent ces éléments de la mythologie antique, l'un à partir de textes primaires, l'autre à partir de l'analyse des publications scientifiques.

L'*Index rerum notabilium* de *Memorata Poetis*

Le projet *Memorata Poetis* (MP) est un projet de recherche intertextuelle (site Internet en cours de finalisation). Il consiste à mettre en lien de courts poèmes (épigrammes épigraphiques et littéraires) composés et transmis de l'Antiquité à nos jours selon des thèmes et des motifs, c'est-à-dire, des schémas récurrents (ex.: La mort, Hercule...). Le corpus contient, à la date de soumission, 15307 textes et s'appuie sur une édition de référence en langue originale (latin, grec, italien anglais et arabe). Il s'agit, au travers de rapprochements textuels et de mises en parallèle, d'identifier l'évolution de l'expression des thèmes et des phénomènes intertextuels qui ont accompagné la transmission, la traduction et la création de ce type de poèmes au travers du temps et des cultures. Ces caractéristiques confèrent au corpus de sources primaires constance et homogénéité, deux critères de sélection importants pour le travail de rapprochement lexicologique.

Ces thèmes sont organisés au sein d'une liste contrôlée et hiérarchisée, l'*index rerum notabilium*, lentement consolidée par plusieurs générations de philologues, selon une arborescence à trois niveaux et six thématiques classiques (Animalia, Arbores et Virentia, Dei et Heroes, Homines, Loca, Res). Il s'agit donc d'explorer une documentation dite primaire et éditée, en amont de la production d'articles scientifiques.

Avec plus de 1300 termes qui le composent ont été traduits, de manière collaborative, une première fois en italien, anglais, allemand, français, grec moderne, avec comme langue de référence le latin en s'en tenant au cadre d'usages disciplinaires de la philologie La traduction entre dans la seconde phase qui consiste à valider ou non des synonymes et des variations linguistiques pour chaque terme.

Le thésaurus PACTOLS de la Fédération et ressources sur l'Antiquité

En aval de ce processus d'élaboration de la connaissance, on retrouve des instruments pour analyser le contenu de la production scientifique. PACTOLS est un thésaurus multilingue (6 langues européennes et l'arabe) développé par la Fédération et Ressources sur l'Antiquité (FRANTIQ), groupement de services du CNRS (Centre national de la recherche scientifique français) http://www.frantiq.fr. Il constitue un outil d'indexation de publications scientifiques spécialisées en archéologie, depuis la Préhistoire jusqu'aux périodes contemporaines et en Sciences de l'Antiquité dans tous les domaines. Il est en particulier exploité dans le Catalogue Collectif Indexé de FRANTIQ (520.000 notices bibliographiques). Le thésaurus est organisé en sept branches (Peuples et Cultures, Anthroponymes, Chronologie, Toponymes, Oeuvres, Lieux, Sujets), elles-mêmes découpées en sous thèmes. Comme dans tout thésaurus, les descripteurs s'inscrivent dans un contexte terminologique précis, spécifié par des relations complexes : hiérarchiques, bien sûr, mais aussi d'équivalences, de synonymie,... Ils sont décrits à l'aide d'une définition et d'une note d'application. Chacun est traduit dans six langues : le français est la langue de référence ; les traductions en allemand, anglais, arabe, espagnol, italien, néerlandais sont mises à jour au fur et à mesure des mises à jour du thésaurus. Certains descripteurs sont illustrés. Enfin, le thésaurus suit la norme ISO 25964 sur l'interopérabilité des thésaurus : chaque terme est identifié par une URI (Unique Resource Identifier) de type ARK (Archival Resource Key - https://fr.wikipedia.org/wiki/Archival_Resource_Key). Cet identifiant lui confère une identité unique dans le web des données et le rend citable. Son format de données est conforme à SKOS (Simple Knowledge Organization Systems) du W3C. Par exemple http://ark.frantiq.fr/ark:/26678/pcrtGYtuX6F7A0 renvoie au concept Némésis et à son environnement sémantique ainsi qu'aux notices liées du Catalogue Collectif Indexé de FRANTIQ.

Ce thésaurus a été créé en 1987 et il est mis à jour en continu par des spécialistes des domaines traités. Il compte aujourd'hui 33.000 descripteurs. Il est utilisé par des bibliothécaires et des documentalistes dans le cadre de catalogues et de bases de données documentaires, comme dans l'enrichissement de bases de données de la recherche. Par exemple, les publications indexées sur la thématique que nous avons retenue dans le cadre de cette présentation sont soit des éditions de textes littéraires antiques, soit des ouvrages avec une riche iconographie : traités sur la sculpture, la mosaïque, les décors architecturaux ou les céramiques…

Dieux et héros de l'Antiquité classique, une thématique commune

On le voit, les deux outils présentés ont été conçus pour répondre à des exigences scientifiques complémentaires qui appartiennent à un même processus global d'élaboration des connaissances. L'organisation des termes comme leurs langues de référence et le multilinguisme des outils et des contenus sont le reflet de l'usage pour lequel ils ont été pensés.

Cependant, des concepts communs se retrouvent, tout au moins sur certaines thématiques. Nous avons choisi de confronter les deux vocabulaires sur les thèmes Dieux et Héros de l'Antiquité classique : comment ces domaines sont-ils traités par l'un, dans le cadre d'analyse du texte antique comme document primaire et par l'autre dans le cadre des publications scientifiques, documents secondaires par excellence ? Comment l'archéologie, l'épigraphie et la littérature élaborent-elles des connaissances sur des sujets similaires ?

Le domaine "Dieux et Héros" de l'*Index rerum notabilium* de *Memorata Poetis* est bien défini: Mythologie, Chrétienne, Musulmane, Religion et Supertitions. Il s'agit d'aligner ce sous-ensemble et ses traductions, avec les branches Anthroponymes/Divinités/Divinités grecques, Divinités romaines, Anthroponymes/Héros/Héros grecs, Héros romains, ainsi que le groupe de termes de Sujets/religion des thésaurus PACTOLS. En effet, les Dieux et Héros de l'Antiquité classique sont regroupés différemment et les termes relatifs à la religion plus généralement sont intégrés à la branche Sujets des PACTOLS, alors que l'ensemble est rassemblé en un seul dossier dans l'*Index* de MP.

La méthode

Dans la mesure où seul le thésaurus PACTOLS est formaté en SKOS avec des identifiants pérennes, nous procéderons à un alignement manuel, terme par terme, avec discussion et ajustement de chacun des systèmes et de son lexique. Les rapprochements stricts ou relatifs des termes, mais aussi l'absence de tel autre dans l'un ou l'autre des outils sont des indicateurs majeurs de positionnement disciplinaires complémentaires qui concernent des moments et des cultures épistémologiques différents. Les théories, méthodes et instruments produits dans le cadre des humanités numériques devraient permettre d'accueillir et de faire converger ces univers terminologiques sur le rapprochement desquels nous travaillons et ce d'autant plus facilement qu'ils en respecteront leurs caractéristiques et les amèneront à dialoguer : exploitation et traitement du réseau de relations entre les termes, confrontation des traductions des vocabulaires, prise en compte, si nécessaire, de l'élaboration d'un métaformalisme commun. Ciotti, 2013, illustre les limites ontologiques rencontrées par les seuls thèmes et motifs. Si ce travail méthodologique de comparaison et de rapprochement entre différents KOS (Knowledge Organization System) ne cherche pas, dans un premier temps à produire un outil numérique, il vise à en préparer l'élaboration.

Par exemple, l'*Index rerum notabilium* contient un seul terme pour identifier Apollon, alors que les PACTOLS comptent neuf façons différentes de qualifier le dieu. L'étude des termes met aussi en avant les écarts contextuels révélateurs des périodes pendant lesquelles ces instruments d'analyse ont été élaborés : par exemple, PACTOLS, bien que régulièrement révisé, gagnerait à être réorganisé pour en faciliter la consultation et mieux suivre l'évolution des connaissances et de la discipline, tout en conservant des descripteurs adaptés aux publications plus anciennes. L'Index s'est construit sur une très lente et longue sélection progressive dont les débuts remontent aux origines de la philologie. Il se caractérise par son unité et sa stabilité mais aussi comme étant un témoin évolutif de la transmission intertextuelle elle-même. Il est destiné à s'enrichir au travers de la mise en parallèle intertextuelle, ce type de recherche scientifique ayant pour but d'identifier des motifs culturels et leurs transformations à l'intérieur des textes.

Bien que les deux instruments se situent sur deux niveaux d'analyse, nombre de concepts se rejoignent et il y a une utilité à pouvoir créer des liens d'un niveau d'analyse à l'autre afin de naviguer transversalement au travers des collections d'informations primaires (ex. textes, ici des épigrammes littéraires et épigraphiques) et secondaires (la littérature scientifique qui en émane). La comparaison des deux instruments permettra aussi leur enrichissement réciproque : les PACTOLS ajouteront les termes latins à ses traductions, l'*Index* bénéficiera d'une troisième phase d'enrichissement de ses traductions, d'une amélioration de sa navigation tout en conservant sa structure traditionnelle. Mais surtout, l'alignement des vocabulaires doit permettre de tirer des conclusions concrètes sur les évolutions que chacun des outils doivent suivre afin de permettre aussi bien au chercheur, mais aussi au grand public de naviguer avec fluidité dans le *mare magnum* documentaire primaire et secondaire en entrant par une même porte conceptuelle.

Résultats attendus

La mise en forme manuelle de ce matériel lexicologique devrait permettre de confronter dans un second temps deux hypothèses. La première est organisationnelle : un système à facettes peut-il permettre une meilleure utilisation de ces connaissances ? L'autre relève de la normalisation : SKOS employé dans le système PACTOLS pourrait-il à terme constituer une évolution nécessaire de l'Index pour finaliser la communication entre les deux systèmes d'indexation ? Les passerelles entre les environnements faciliteront ainsi le passage de l'un à l'autre, au profit des usagers dans la mesure où l'utilisation de termes identiques pour un même concept crée des ponts et des ouvertures à la connaissance en favorisant le dialogue entre les disciplines (archéologie, épigraphie, philologie). La thématique choisie pour cet exercice est bien connue aussi du grand public et fait partie de l'héritage culturel des sociétés gréco-romaines antiques. Pour favoriser encore la circulation des connaissances, nous compléterons les articles Wikipédia avec les termes des vocabulaires pour favoriser l'accès aux ressources primaires (les textes anciens utilisé par *Memorata Poetis*) ou secondaires (les publications scientifiques indexées avec PACTOLS). Il s'agit donc d'ajouter des liens d'autorité externes à la fin de ces articles de l'encyclopédie.

Les résultats de cette analyse appliquée, via l'alignement, au sous-ensemble des dieux et héros de la Mythologie doit permettre d'étendre le travail à l'ensemble des deux vocabulaires, avec un reversement et un enrichissement réciproque des termes traduits dans plusieurs langues européennes et du pourtour du bassin méditerranéen.

Le thésaurus PACTOLS (http://pactols.frantiq.fr) est utilisé par 40 bibliothèques de la Fédération et ressources sur l'Antiquité et par la Très Grande Infrastructure de Recherche française Huma-Num dans son moteur de recherche ISIDORE. Il est l'une des contributions françaises à l'infrastructure européenne DARIAH. Il est aussi impliqué dans des projets européens comme MULTITA, piloté par les Musées royaux de Belgique et dans ARIADNE, infrastructure européenne pour la recherche archéologique.

Le projet Memorata Poetis (financement national PRIN 2010/11 - http://www.memoratapoetis.it/public en cours de finalisation) est dirigé par le Professeur Paolo Mastandrea et implique huit unités universitaires ainsi que de recherche: les Universités de Venise "Ca' Coscari", de Cagliari, de la Calabre, de Padoue, de Pérouse, de Rome "La Sapienza", de Rome "Tor Vergata", et l'Istituto di Linguistica Computazionale "A. Zampolli" du Consiglio Nazionale delle Ricerche (CNR) à Pise.

Bibliography

Babik, W. and International Society for Knowledge Organization. Polish Chapter (eds). (2014). *Knowledge Organization in the 21st Century : Between Historical Patterns and Future Prospects ; Proceedings of the Thirteenth International ISKO Conference 19 - 22 May 2014, Kraków, Poland*. (Advances in Knowledge Organization 14). Würzburg: Ergon.

Borillo, M. (1984). *Informatique pour les sciences de l'homme: limites de la formalisation du raisonnement*. (Philosophie et langage). Bruxelles: P. Mardaga.

Burr, E. (ed). (1994). *The Chiron Dictionary of Greek & Roman Mythology: Gods and Goddesses, Heroes, Places, and Events of Antiquity*. Wilmette, Ill.: Chiron.

Caputo, M. and Lamé, M. (En cours). Humanités numériques, organisation de la connaissance et convergences disciplinaires : quelques études de cas en sciences de l'Antiquité. In International Society of Knowledge Organization (ISKO) (ed), *Systèmes D'organisation Des Connaissances et Humanités Numériques, International Society of Knowledge Organization (ISKO) Conference, Strasbourg, 2015*. Strasbourg.

Cazanove, O. de, Scheid, J. and Collège de France (eds). (2003). *Sanctuaires et sources dans l'antiquité: les sources documentaires et leurs limites dans la description des lieux de culte*. (Collection du Centre Jean Bérard 22). Napoli, Italie: Centre Jean Bérard.

Ciotti, F. (2014). Tematologia e metodi digitali: dal markup alle ontologie. In Alfonzetti, B., Baldassari, G. and Tomasi, F. (eds), *I cantieri dell'italianistica. Ricerca, didattica e organizzazione agli inizi del XXI secolo. Atti del XVII congresso dell'ADI – Associazione degli Italianisti (Roma Sapienza, 18-21 settembre 2013)*. Roma: Adi.

Fondation pour le lexicon iconographicum mythologiae classicae (ed). (1981). *Lexicon iconographicum mythologiae classicae, LIMC*. 16 vols. Zürich, Suisse: Artemis Verlag.

Fondation pour le lexicon iconographicum mythologiae classicae and J. Paul Getty museum (eds). (2004). *Thesaurus cultus et rituum antiquorum (ThesCRA)*. 7 vols. Basel; Los Angeles: Fondation pour le Lexicon iconographicum mythologiae classicae ; J. Paul Getty museum.

Gnoli, C. (2012). Metadata About What?. *Knowledge Organization*, **39**(4): 268–75.

Gnoli, C. (2014). Boundaries and overlaps of disciplines in Bloch's methodology of historical knowledge. In Babik, W. (ed), *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects : Proceedings of the Thirteenth International ISKO Conference, 19-22 May 2014, Kraków, Poland*. Würzburg: Ergon, pp. 129–35.

Gnoli, C., Rosati, L. and Marino, V. (2006). *Organizzare la conoscenza : dalle biblioteche all'architettura dell'informazione per il web*. (Internet e..). Milano: Hops Tecniche nuove.

Institut Thesaurus Linguae Latinae (2009). *Thesaurus Linguae Latinae (TLL) Online*. De Gruyter.

Ogden, D. (ed). (2007). *A companion to Greek religion*. (Blackwell companions to the ancient world). Malden; Oxford; Victoria: Blackwell.

Otlet, P., Peeters, B., Fayet-Scribe, S. and Wright, A. (2015). *Le livre sur le livre: traité de documentation*. (Réflexions faites). Bruxelles: Les Impressions nouvelles.

Papadopoulou, T. (1999). Literary Theory and Terminology in the Greek Tragic Sholia: the Case of ΠΛΑΣΜΑ*. *Bulletin of the Institute of Classical Studies*, **43**(1): 203–10. doi:10.1111/j.2041-5370.1999.tb00489.x.

Pierazzo, E. (2015). *Digital scholarly editing: theories, models and methods*. Farnham; Burlington: Ashgate.

Talbert, C. H. (1975). The Concept of Immortals in Mediterranean Antiquity. *Journal of Biblical Literature*, **94**(3): 419–36. doi:10.2307/3265162.

Harvesting History: Democratizing The Past Through The Digitization Of Community History

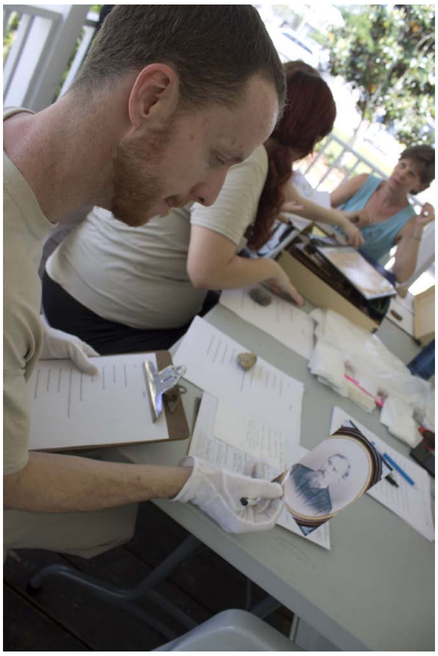
Connie Lee Lester
Connie.Lester@ucf.edu
University of Central Florida, United States of America

In 1993-94 Roy Rosenzweig and David Thelen conducted a set of surveys that resulted in the publication of a seminal work titled *The Presence of the Past*. The team of historians, museum curators and students surveyed the American public about their interaction with history in order to understand "better ways of connecting academic historians with larger audiences" (Rosenzweig and Thelen, 1998). Their study concluded that Americans' interest in history related to those events and groups that intersected with their own lives. They universally trusted museums because they could see the objects and images that were the "stuff" of history, but family stories, letters, and photographs defined history for most of the survey participants. With the publication of the survey, Rosenzweig and Thelen entered a historiographic discussion that centered on ways to incorporate history from the "bottom up" and to make the museum experience more interactive (Wallace, 1996; Hayden, 1997; Gardner and LaPaglia, 2004; Lewis, 2005; Gordon, 2010; Anderson, 2012; and Falk and Dierking, 2012).

Almost twenty years after Rosenzweig and Thelen conducted their survey, William G. Thomas, Patrick D. Jones and Andrew Witmer "advanced a movement to democratize and open history" with the founding of the History Harvest Project at the University of Nebraska (Thomas, Jones, and Witmer, 2013). Incorporating their own digital history experience into the classroom, Thomas, Jones, and Witmer developed student projects that operated at the "intersection of digital history and experiential learning." Beginning in 2010 and continuing to the present, students worked with specific communities to organize History Harvest events at which they photographed objects and scanned images offered by individuals from their personal collections. The photographs and scans were uploaded into an Omeka database and organized into collections and digital exhibits. The project's aim is to "make invisible archives and stories more visible." The founders argue that to be successful, a History Harvest must be "organic, grassroots, and local."

At the University of Central Florida (UCF), a Regional Initiative for Collecting History, Experiences, and Stories (RICHES™) Project has advanced the concept of student-public engagement through the use of History Harvests to collect "invisible archives and stories" by making the presentation of the digitized material interactive and democratizing the ability of site users to utilize the data in the collections to develop their own narratives. This transformational addition to the efforts to democratize history provides the user with the data, the tools to analyze the data, and the digital space to organize the data in order to create a narrative.

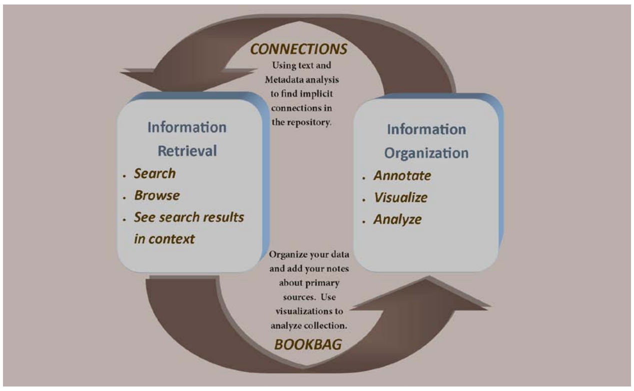
Working in Florida, students have access to a history that celebrates more than 500 years of interaction between Native Americans, Europeans, and Africans. The geographical position of Florida with its extensive coastline means that the history of the area has always been global. Interaction between the people of Caribbean Islands and those living in Florida pre-dated European claims for the region and expands the scope of scholarship. Successive waves of Spanish, French, British and American settlers altered the landscape and created new borderlands that continued until well into the twentieth century. Currently, the third most populous state in the nation, with an economy that ranges from agriculture to tourism to space exploration and high tech industries, Florida's history is deep, diverse, and largely unexplored.



Oviedo History Harvest, 2015

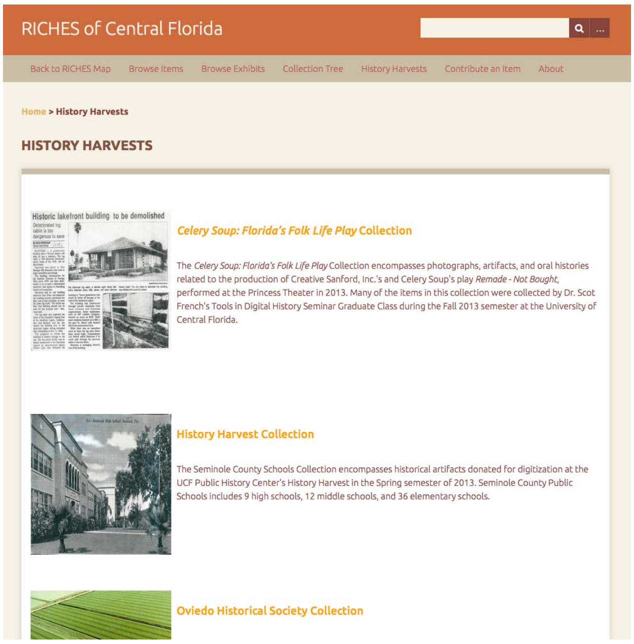
Students in graduate classes in Public History at UCF have completed three History Harvests and three additional Harvests are planned for 2016. The largest of the completed Harvests was accomplished through a partnership between the Introduction to Public History class, the Oviedo (Florida) Historical Society and a RICHES partnership tech firm, EZ-Photo-I/O Track. The students worked with the historical society to understand the history of the 137-year-old agricultural community that was transformed by the founding of UCF and the influx of thousands of students and faculty only three miles from the center of town. Based on the information provided by the historical society and a National Endowment for the Humanities-funded, society-published book (Adicks and Neely) the students conducted sixteen oral histories with descendants of the original settlers. In order to generate community awareness of the upcoming History Harvest students participated in public events and obtained permission to post information on the community and historical society social media sites. Building on the planning documents produced by previous student-led History Harvests, they organized the release forms, prepared to obtain the stories associated with items brought for scanning, and worked with EZ-Photo staff to facilitate the scanning process. More than 500 images, documents, three dimensional items, pamphlets, and scrapbooks were scanned on the day of the History Harvest. They told the story of citrus grove owners, farm laborers and fruit packers, migrants and immigrants to the community, schools, churches, and community social life.

The UCF/RICHES History Harvest pushed the use of the collected data to a new, level by placing the material in RICHES Mosaic Interface™ https://richesmi.cah.ucf.edu . RICHES MI was initially created with the help of a National Endowment for the Humanities Startup Grant, and is an Omeka archive with multiple custom plugins and a front end that is a unique mechanism for searching the archive. RICHES MI users can search the database using natural language, tags, and topics, and browse by collection categories to maximize their search results; analyze the results of their search using the RICHES-developed "Connections" feature to show the relationship between the returned item and other items in the database; and visualize their results on a Google map and through digital exhibits, map overlays, and other visualizations. Users can save their search results in a Bookbag where they can annotate individual items, store items in folders, map the items collected in folders on a Google map and a timeline, and analyze the connections between items in the folders. They can also develop a narrative using the story board feature to organize their data. Finally, a search in the Omeka archive provides the user with links to digitized primary and secondary sources relevant to the individual object as well as links to other databases or websites that might be useful to the researcher.


CONNECTIONS
Using text and Metadata analysis to find implicit connections in the repository.
Information Retrieval
. Search
. Browse
. See search results in context
Information Organization
. Annotate
. Visualize
. Analyze
Organize your data and add your notes about primary sources. Use visualizations to analyze collection.
BOOKBAG

Schematic of RICHES Functionality

Student Harvests, like their counterparts elsewhere, also create digital exhibits to organize the collected images and documents into an interpretative narrative. Faculty and students gain much through their participation in the History Harvests. Students acquire digital skills in metadata writing, archiving and exhibit creation; they learn to conduct oral histories and plan events; and they develop an understanding of the connection between local history and national/global issues. Each student's work is cited by name in the metadata and in digital exhibits. Faculty participants are organizing larger projects (i.e. Parramore Project and Glass Bank Project) that include History Harvests and are expected to result in scholarly and pedagogical publications.



HISTORY HARVESTS

Celery Soup: Florida's Folk Life Play Collection

The *Celery Soup: Florida's Folk Life Play* Collection encompasses photographs, artifacts, and oral histories related to the production of Creative Sanford, Inc.'s and Celery Soup's play *Remade - Not Bought,* performed at the Princess Theater in 2013. Many of the items in this collection were collected by Dr. Scot French's Tools in Digital History Seminar Graduate Class during the Fall 2013 semester at the University of Central Florida.

History Harvest Collection

The Seminole County Schools Collection encompasses historical artifacts donated for digitization at the UCF Public History Center's History Harvest in the Spring semester of 2013. Seminole County Public Schools includes 9 high schools, 12 middle schools, and 36 elementary schools.

Oviedo Historical Society Collection

History Harvests in Omeka

The next three projects, which are funded by a National Endowment for the Humanities "The Common Good: The Humanities in the Public Square" grant, will push the boundaries of History Harvests further: students will be using the oral histories and the collections obtained through History Harvests to produce a documentary film on an African American community's struggles with urban development, to understand a coastal community's relationship to an iconic building, and to provide evidence for two MA thesis projects dealing with Orlando's LGBTQ community. The African American community project pairs students in an undergraduate history class with students at a predominantly black high school to explore changes that the construction of public buildings, entertainment and sports venues, an interstate highway, and the proposed expansion of UCF to a downtown campus have had on what was once a vibrant and cohesive African American community. The material collected will be used by a second undergraduate Honors class to produce a documentary film. This will be the fourth student produced film supervised by the course faculty. The second History Harvest will focus on an iconic building, the so-called Glass Bank, in Cocoa Beach, Florida. Constructed in the 1960s, the Glass Bank was demolished in 2014, but not before a partnership between RICHES and the Institute for Simulation and Training produced a 3D scan of the building. Recreated as it existed in 1963, the Glass Bank will serve as a vehicle for collecting images, artifacts and stories about the building as a center of community activity and economic development. See the video of the 3D scan https://www.youtube.com/watch?v=DtVETzCND4M Finally, students will work with the GLBT Virtual Museum and the Orange County Regional History Center to harvest images, oral histories, and artifacts that tell the story of the gay destination, Parliament House, and the effect of the HIV/AIDS epidemic in Orlando.

The RICHES project builds upon the insights of Public History scholarship and Digital scholarship to understand the complex past of local communities and to connect those histories to larger narratives. Using History Harvests to build a bridge between individuals, families, and local museums and historical societies and the academic community of students and faculty provides a service to the public and experiential learning for students. Awareness of the interaction between local and the national/global events enables students to consider their own scholarship in new ways. Placing the material harvested in an interactive database like RICHES Mosaic Interface democratizes the history collected by providing context, analytical tools, and space for organizing personal narratives.

Bibliography

Wallace, M. (1996). *Mickey Mouse History and Other Essays on American Memory*. Philadelphia, PA.

Hayden, D. (1997). *The Power of Place: Urban Landscapes as Public History*. Cambridge, MA.

Adicks, R. and Donna N. (1998). *Oviedo: Biography of a Town*. Ovideo, FL.

Rosenzweig, R. and Thelen, D. (1998). *The Presence of the Past: Popular Uses of History in American Life*. New York.

Gardner, J. B. and LaPaglia, P. S. (eds). (1999), (2004). *Public Essays from the Field*. Malabar, FL.

Lewis, C. (2005). *The Changing Face of Public History: The Chicago Historical Society and the Transformation of an American Museum*. DeKalb, IL.

Gordon, T. S. (2010). *Private History in Public: Exhibition and the Settings of Everyday Life*. Nanham, MD.

Anderson, G. (ed). (2012). *Reinventing the Museum: The Evolving Conversation on the Paradigm Shift*. Lanham, MD.

Falk, J. H. and Dierking, L. D. (2012). *The Museum Experience Revisited*. Walnut Creek, CA.

Thomas, W. G., Jones, P. D. and Witmer, A. (2013). "History Harvests: What Happens When Students Collect and Digitize the People's History?" *Perspectives on History: The Newsmagazine of the American Historical Association* (January 2013) https://www.historians.org/publications-and-directories/perspectives-on-history/january-2013/history-harvests (accessed October 15, 2015).

If Paintings were Plants: Measuring Genre Diversity in Seventeenth-Century Dutch Painting and Printmaking

Matthew Lincoln
mlincol1@umd.edu
University of Maryland, United States of America

Tracing the "origins of pictorial species" (to borrow Larry Silver's turn of phrase) has long been an interest of art historians (Silver, 2006). The emergence of distinct genres of painting (e.g. dedicated landscapes or still-lifes) in the sixteenth and seventeenth centuries in Europe, and in the Netherlands in particular, has proven especially fascinating. Historians of art and economics have hypothesized that, by specializing in standalone still-lifes, landscapes, or so-called "genre scenes" of everyday life, painters may have reaped two advantages: an opportunity to distinguish themselves in the uniquely competitive art market in the sixteenth- and seventeenth-century Netherlands; and the ability to efficiently paint similar compositions over and over again (Chong, 1987; Montias, 1988). But would professional printmakers also have adopted this specialization strategy? Or did the medium, which was often put to use making reproductions after other artists' designs, instead favor etchers and engravers willing to render the works of a wide variety of artists? Existing case studies present conflicting evidence. How can we test this question at scale?

While Silver only invokes speciation as a metaphor, ecology may offer a useful quantitative model for thinking about genre specialization. A common measurement of species diversity (Shannon's diversity metric) can be used to characterize artists' relative specialization or diversification in genre, thus allowing us to gain a broader perspective on printmakers' specialization or diversification strategies. I will first demonstrate how this index can detect Dutch painters' trend towards genre specialization from a database of paintings seventeenth-century Dutch household inventories, and a comprehensive database of Dutch paintings in modern-day museum collections. I will then use it to test whether or not we can detect similar results in a database of prints maintained by the Rijksmuseum.

Methodology

Whether looking at the diversity of species within an ecosystem, or the variety of different industries within a state, diversity measures have to account for two dimensions:

1. Categorical: How many discrete classes are observed?
2. Allocation: How even is the distribution of units among categories?

Shannon's measurement of diversity (D_s), a widely-used metric, captures both of these dimensions of diversity.[1] Originally developed to characterize entropy in information transmission, this metric of diversity has been applied to the studies as diverse as ecological diversity, economic specialization, and racial segregation (Gibbs and Martin, 1962; Ottaviano et al., 2003). To measure whether specialization or generalization was more favored by painters and printmakers, each artist's oeuvre is treated as a "population" with a single diversity score calculated per artist. By this measure, a population whose members are distributed evenly across several different species/categories will have a higher diversity index than a population whose members are largely concentrated in just one category.

Data

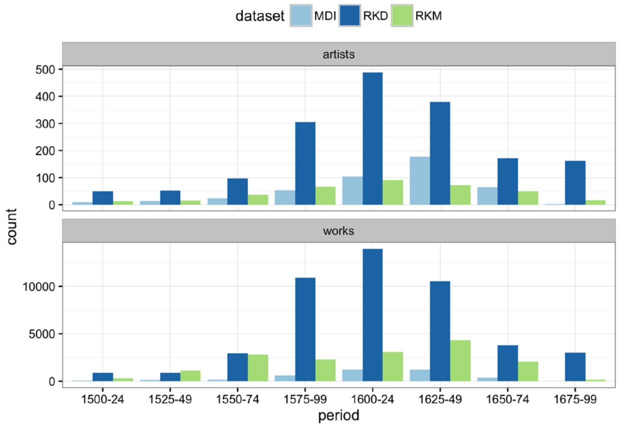


Figure The number of unique artists and artworks represented in each dataset, subdivided by birth year

This study is based on two sources of information about paintings, and one source for prints.

The first, a modern resource, is the *RKDimages* database compiled by the Rijksbureau voor Kunsthistorische Documentatie.[2] This catalog of Dutch and Flemish artworks extant in collections around the world contains approximately 13,000 dated and attributed paintings that have each been tagged with a series of keywords (on average between 6 to 7 keywords per painting) describing their subject matter.[3] The scale of the RKD database makes it unfeasible for the individual researcher to manually categorize each artwork into a single broad subject category. Therefore, I identified clusters of artworks that shared groups of keywords though community detection on a constructed graph where each object was connected to others based on shared RKD subject keywords.[4] I then checked the resulting groups manually to confirm that they did, in fact, corresponded relatively well to common genre categories. The resulting groups roughly encompassed: 1) portraits, 2) still lifes, 3) landscapes, 4) religious paintings, and 5) a looser array of other works that featured multiple figures (generally genre scenes or history subjects).

Because the surviving paintings in the RKD database are a biased proxy of the **actual** patterns of paintings produced in the seventeenth century, it is crucial to compare the trends derived from the RKD's modern database against contemporary archival records. The Montias Database of 17th Century Dutch Art Inventories, maintained by the Frick Art Reference Library, contains information on household inventories from Amsterdam that were recorded between 1575 and 1700.[5] Of these inventories, 1153 contained at least two paintings The Montias Database has 86 different subject headings, which have been manually grouped into the same general set of subject headings that we used for the RKD database.[6] The MDI describes 34,147 paintings, of which 26,349 (about 77% of the total) have an identified subject (the rest are labeled "unknown"), with 4,377 of those described paintings (about 13% of the total) attributed to a specific artist.[7] The Montias inventories are also an imperfect reflection of seventeenth-century painting production, being biased towards rich collectors, mostly in Amsterdam, who died with outstanding debt. However, if both the modern and contemporary datasets reveal similar patterns in specialization, this would strengthen the case for claiming that a trend towards specialization existed historically (De Vries, 1991:259-260.)

Results

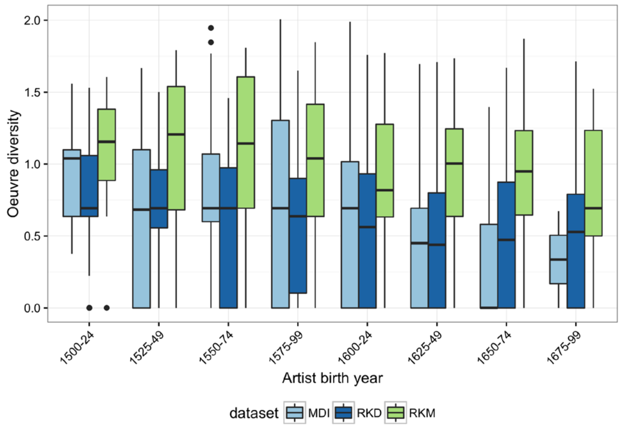


Figure 1. The oeuvre diversity ranges of painters (Montias and RKD datasets) and printmakers (RKM dataset) born at different points between 1500 and 1700

Unfortunately, there are virtually no seventeenth-century inventories that catalog individual prints. Instead, we rely solely on the surviving prints in the collection of the Rijksmuseum[8]: an imperfect source, though one that is also unparalleled in its coverage of known surviving prints from this period. The Rijksmuseum has classified their artworks based on the ICONCLASS system for tagging iconography in European art[9], and this has also been mapped to the same broader categories used in the Montias database. Multiple impressions of the same print have been roughly disambiguated by removing prints with a duplicate engraver, title, and dates. This study is also only considering reproductive prints, so prints made by engravers or etchers after their own designs are excluded from this analysis.

We find that, although both the Montias and RKD paintings datasets show wide variation, with both highly specialized painters and highly diversified ones, both datasets reflect an increasing number of specialized painters born after 1600, as shown by a decreasing median oeuvre diversity. On the other hand, the median oeuvre diversity of printmakers in the Rijksmuseum dataset remains consistent during the entire period of study. This confirms the widely-held hypothesis that an increasing number of Dutch painters defined a niche for themselves by specializing in a particular genre. The results also appear to support the previously-unexamined hypothesis that reproductive printmakers instead favored making prints after a wide array of artworks; printmakers who did define highly specialized niches appear to have been the exception, rather than the norm.

So what subjects did these specialists prefer? Prolific specialized painters overwhelmingly favored landscapes: of those artists in the bottom diversity quartile (i.e. the 25% most specialized painters in the Montias database), almost 85% of their paintings are described as landscapes, followed in a distant second by still-life paintings. Landscape was a genre that was both highly conventionalized - it was easy to produce endless variations on the same general set of topographical motifs - and also amenable to a very efficient technique - a landscape could be rendered in broad brushwork with a limited palette and still be an aesthetic success (Goedde, 1997). Still-life paintings in this period were also, by in large, the purview of specialists. They may comprise a much smaller share of the total number of paintings in the Montias database because, unlike landscapes, the aesthetic effect of still lifes was often dependent on the painter's mimetic skill and illusionistc finish - not a technique conducive to speedy production.

On the other hand, of those printmakers who **did** specialize in particular genres of prints, we do not find a single dominant theme. Rather, a few specialties rise to the top: "news" prints depicting current events, architectural illustrations, and allegorical or biblical series prints. Those printmakers who did specialize (Table 1) did not rely on prints as their main means of support; many were specialist painters who happened to produce prints as well. But these printmakers were outnumbered by those professional printmakers (Table 2) who were willing and able to render reproductions after a wide variety of artists. This flexibility could have presented an attractive insurance policy for print publishers, who had to continually react to the demands of a quickly-moving market for artistic prints and illustrations, while also appealing to the seventeenth-century function of prints as encyclopedic sources of knowledge (MacGregor, 1999:395).

artist name	works	div	subjects
Abraham Dircksz Santvoort	123	0.74	topographical views, history prints
Allaert van Everdingen	113	0.77	landscape, animals
Adriaen van Ostade	85	0.67	genre, low-life
Isaac Vincentsz van der Vinne	82	0.23	heraldry
Reinier Nooms	72	0.58	seascape
Cornelis Dusart	57	0.76	portraiture
Theodoor van Thulden	52	0.68	antiquity, mythology
Cornelis Pietersz Bega	36	0.62	genre
Anthonie Waterloo	35	0.67	landscape

Table 1: The most productive Dutch and Flemish specialist printmakers (those falling below the 45th diversity percentile). Note that the count of "works" treats print series as a single work

artist name	works	div	subjects
Jan Luyken	2,047	1.64	bible scenes, seascape, genre, historical, architecture, titlepages
Caspar Luyken	454	1.74	bible, landscape, historical, genre, maps
Aegidius Sadeler	238	1.80	landscape, portraiture, allegory, mythology, religious
Jacob Matham	229	1.73	allegory, mythology, portraiture, biblical
Hendrick Goltzius	228	1.80	biblical, portraiture, allegory, antiquity, mythology, landscape
Crispin van de Passe (I)	203	1.80	moralizing allegories, portraiture, devotional, botanical, biblical
Johannes Wierix	215	1.54	portraiture, biblical, allegory, genre, mythology, devotional,
Abraham Bloteling	188	1.72	landscape, genre scenes, portraiture, mythological

Cornelis Bloemaert (II)	180	1.69	saints & other religious, biblical, portraiture
Raphaël Sadeler (I)	156	1.57	devotional series, biblical, allegory, mythology, titlepages

Table 2: The most generalist Dutch and Flemish printmakers (those falling above the 85th diversity percentile)

Bibliography

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10): P10008 doi:10.1088/1742-5468/2008/10/P10008.

Chong, A. (1987). The Market for Landscape Painting in Seventeenth-Century Holland. In Sutton, P. C. and Blankert, A. (eds), *Masters of Seventeenth-Century Dutch Landscape Painting*. Boston: Museum of Fine Arts, pp. 104–20.

Csardi, G. and Nepusz, T. (2006). The igraph Software Package for Complex Network Research. *InterJournal*: 1695. http://igraph.org.

De Vries, J. (1991). Art History. In Freedberg, D. and De Vries, J. (eds), *Art in History, History in Art: Studies in Seventeenth-Century Dutch Culture*. Santa Monica: Getty Center for the History of Art & the Humanities, pp. 249–71.

Gibbs, J. P. and Martin, W. T. (1962). Urbanization, Technology, and the Division of Labor: International Patterns. *American Sociological Review*, **27**(5): 667–77. doi:10.2307/2089624.

Goedde, L. O. (1997). Naturalism as Convention: Subject, Style, and Artistic Self-Consciousness in Dutch Landscape. In Franits, W. E. (ed), *Looking at Seventeenth-Century Dutch Art: Realism Reconsidered*. Cambridge: Cambridge University Press, pp. 129–43.

MacGregor, W. B. (1999). The Authority of Prints: An Early Modern Perspective. *Art History*, **22**(3): 389–420. doi:10.1111/1467-8365.00163.

Montias, J. M. (1988). Art Dealers in the Seventeenth-Century Netherlands. *Simiolus: Netherlands Quarterly for the History of Art*, **18**(4): 244–56. doi:10.2307/3780702.

Montias, J. M. (2015). The Montias Database of 17th Century Dutch Art Inventories. Frick art reference library Database. http://research.frick.org/montias/home.php.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Sipson, G. L., Solymos, P., Stevens, M. H. H. and Wagner, H. (2015). Vegan: Community Ecology Package R Package (version 2.2-1) R Package (version 2.2-1), ms http://cran.r-project.org/package=vegan.

Ottaviano, G. I. P., Pinelli, D., Maignan, C. J. and Rullani, F. (2003). *Bio-Ecological Diversity Vs. Socio-Economic Diversity: A Comparison of Existing Measures*. SSRN Scholarly Paper. http://papers.ssrn.com/abstract=389043.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Silver, L. (2006). *Peasant Scenes and Landscapes: The Rise of Pictorial Genres in the Antwerp Art Market*. Philadelphia: University of Pennsylvania Press.

Woude, A. M. van der (1991). The Volume and Value of Paintings in Holland at the Time of the Dutch Republic. In Freedberg, D. and De Vries, J. (eds), *Art in History, History in Art: Studies in Seventeenth-Century Dutch Culture*. Santa Monica: Getty Center for the History of Art & the Humanities, pp. 285–329.

Notes

[1] Shannon's diversity index D_s is defined as the negative sum of the proportions of every class size within the population multiplied by their logged equivalents, where n_i is the number of observations in class i, and N is the total population size: D_s=-$\Sigma n_i N l n n_i N$ For the original derivation of Shannon's diversity, see (Shannon and Weaver, 1949); for the R implementation of this formula, see (Oksanen et al., 2015).

[2] https://rkd.nl/en/explore/images

[3] These keywords have been assigned by hand by researchers at the RKD, building on the index cards typewritten by Hofstede de Groot (1863-1930) that have served as the foundation of the RKD's current digital databases.

[4] On the community detection algorithm, see (Blondel et al., 2008); implemented in R by (Csardi and Nepusz, 2006).

[5] http://research.frick.org/montias/home.php

[6] The number of subject headings detailed by Montias are small enough that it was feasible to manually generate a concordance between the 86 original subject headings and the ten subject headings used by Van der Woude in his study of the same database: "old testament", "new testament", "other religious", "mythology-allegory", "history", "landscape", "genre", "still life", "portrait", "animals", "other", and "unknown"; (van der Woude, 1991).

[7] This limited level of description common in collection inventories from the seventeenth century. While it was common to describe the subject of the painting and its size, notaries generally did not make an attribution of an artwork unless its painter was well-known enough that its attribution would have impacted the painting's monetary value.

[8] https://www.rijksmuseum.nl/

[9] http://iconclass.org/

Quantitative Analyses of Chinese Poetry of Tang and Song Dynasties: Using Changing Colors and Innovative Terms as Examples

Chao-Lin Liu
chaolinliu@gmail.com
National Chengchi University, Taiwan, Republic of China

Introduction[1]

Tang (618-907 AD) and Song (960-1279) dynasties are two very important periods in the development of Chinese literary. The majority forms of the poetry in Tang and Song were Shi (詩) and Ci (詞), respectively. Tang Shi and Song Ci established crucial foundations of the Chinese literature, and their influences in both literary works and daily lives of the Chinese communities last until today.

Recognizing the importance of Tang Shi, a Chinese emperor of the Qing dynasty (1644-1912), Kangxi, ordered to compile a collection of Tang poems, *Quan-Tang-Shi* (*QTS*, 全唐詩). *QTS* contains nearly 50 thousand works of about 2200 poets. A similar effort for compiling a collection of Song Ci from the private sector began in the Ming dynasty (1368-1644), and achieved a collection called *Quan-Song-Ci* (*QSC*, 全宋詞) in the early Republican period of China (ca. 1937). *QSC* contains around 20 thousand works of about 1330 poets. The exact statistics about *QTS* and *QSC* may vary slightly depending on the sources.

In the past more than a thousand years, literary and linguistic researchers have had done a myriad of research about the poetry of the Tang and Song dynasties. Hence, it is beyond our capacity and not our objective to review the literature in this abstract. Traditional researchers studied and compared poetic works that were produced by different authors and in different time periods to produce insightful and invaluable analyses and commentaries (Tao, 1999). Most of the time, the researchers focused on the poems of selected poets. Even when computing supports become available, studying poems of specific poets (Jiang, 2003) is still an important and popular type of research in poetry.

Software tools facilitate the analysis of poetry from a panoramic perspective, and may lead to applications that would be very challenging in the past. For instance, Zhou and his colleagues (2009) analyzed the contents of collected couplets and Tang poems for creating couplets. Yan and his colleagues (2013) considered topic modeling in automatic composition of Chinese poetry. Lee (2012) concentrated on the linguistic analysis and teaching of *QTS*.

In this presentation, we will discuss some interesting findings in some quantitative analyses of *QTS* and *QSC*.

Colors and Imageries

Colors are an important ingredient in everyday lives, and actually carried important meanings in religion and social statuses in pre-modern Chinese societies. Wong (2011), a specialist in colors, discussed hidden meanings of colors in China in his book on “The colors of China”.

The beauty and imageries conveyed in the poetry originate from the collocations of the written words in the poems. Lo (2008) and Huang (2004) attempted to classify terms in the poetry by their semantic categories. The results can then serve as a foundation for analyzing the imageries hidden in the poems. Liu and his colleagues (2015) emphasized that colors play a crucial role in painting the imaginaries of poems: “colors in poems are like audios in movies”, and they analyzed the words related to colors in *QTS*.

Using methods for text analysis, one may analyze occurrences and collocations of colors in *QTS* and *QCS*. The main contribution of our work will be illustrating meaningful applications of text analysis for linguistics and literary. The collocations of color words that appeared frequently in *QTS* are certainly interesting (Liu, 2015). Yet, the analysis can be extended in at least two directions. First, did a poet have specific preferences on some collocations? Second, how were the collocations used by different authors?

Bai Juyi[2] has the largest number of works in *QTS*. He used “白髮”[3] with “青衫”[3], “青雲”[3], “丹砂”[3], and “青山”[3], and “白首”[3] with “青山” and “紅塵”[3] relatively often. Liu Changqin[2] , another important poet in the mid Tang period, used “白髮” with “滄洲”[3], and “白首” with “青山”, “滄洲”, and “青春”[3] relatively often. These different words and collocations convey the imagery of “aging”, and the variations in the word choices shed light on the subtle differences between the poets about how they expressed emotions about aging.

“白雲”[4] is a very frequent word in *QTS*. Collocating with different words would create different imageries in the poems, e.g., “黃葉”[5], “滄海”[6], “清露”[7], and “流水”[8]. Each of these collocations may brew a different scene in readers’ minds, and some of these collocations are more popular than others. It should be interesting for researchers to extract the source poems[9] from the *QTS* to thoroughly study them.

Liu et al. (2015) reported that white (“白”) is the most frequent color in *QTS*. We may check and find that red (“紅”) is the most frequent color in *QCS*. It is interesting to investigate the changes (and their causes) of the popular colors from *QTS* to *QCS*. Again, using text analysis methods, we can find a good approximation of the trend, though obtaining the precise frequencies of the colors requires the techniques of word sense disambiguation (WSD). For instance, “金”[10] could represent a material or a color, so WSD is necessary to achieve precise statistics.

In *QSC*, the most frequent six colors are in the order of “紅”, “青”, “黃”, “綠”, “白”, and “碧”[11], while, in *QTS*, the most frequent six colors are “白”, “青”, “紅”, “黃”, “碧”, and “綠”. “紅塵”, “殘紅”, “紅妝”, “紅葉”, “紅袖”, “紅日”, and “紅樓”[12] are some of the most frequent red words in *QSC*. The changes in the dominant colors from *QTS* to *QSC* may be a result of the selection process and may be a result of the cultural shift, and is an academically interesting issue to purse further.

Word Inventions and Influences

Liu and Wang (2012) proposed a method to measure and compare the influences of the poems of a poet. Their methods considered whether or not the poems were selected to be included in famous collections.

While our goal is not to challenge Liu and Wang's viewpoint, we would propose to consider also whether poets created new words that were used by later Chinese generations. Isn't it practically meaningful and academically significant to create new words that future generations continue to use? At the time of Tang and Song, poets were at an excellent stage of the Chinese history to achieve such a cultural impact.

We conduct an analysis of frequent bigrams in *QTS* and *QSC*, and compare the differences. Words that appeared only in *QSC* are candidates of new words which were invented in Song dynasty. Words that appeared only in *QTS* are candidates of words that failed to survive in Chinese language. Although this process does not really guarantee a water-proof theoretical foundation for word invention, the findings should still serve as a persuasive factor in linguistic, literary, and historical research.

Here are some of such findings. “紅塵”[13] appeared in both *QTS* and *QSC*. “惺忪”[13] is a word that appeared in *QSC*[14] but not in *QTS*, and is still being used in modern Chinese. “空門”[13] is a word that appeared in *QTS* but not in *QSC*, and is being used in Chinese. “武皇”[13] is a word that appeared in *QTS* but not in *QSC*, and is not normally used in Chinese. “酴醾”[15] represents another type of instance. This word appeared much often in *QSC* than in *QTS*.

Discussions

A major challenge in analyzing the words in Chinese poetry is word segmentation. Traditional experience indicates that most words in poetry consist of one or two characters (2005). Relying on this heuristic, we can algorithmically analyze the corpora containing about 4.9 million characters at least approximately. To really understand and appreciate the poetry, one should not read them verbatim. Metaphor recognition can be essential for revealing the real intentions of the poets.

Observations resulting from quantitative analysis of *QTS* and *QSC* open windows to promising research opportunities about *QTS*, *QSC*, and their transition. In addition to colors, terms about astronomical objects, floral entities, meteorological phenomena, and geographical sights, are important participants in poetry. Innovative collocations of them paint impressive imageries in readers' minds. Good computational tools can help researchers explore poets' worlds more efficiently.

Responses to the Reviews

We are thankful for the precious critiques and comments of the DH2016 reviewers. Our responses, which could not exceed 300 words, for the comments cannot fit into this final version because of length constraints, so we place our complete responses in a separate file online at <http://www.cs.nccu.edu.tw/~chaolin/papers/dh2016liu.responses.pdf>.

Bibliography

Huang, Ch.-R. (黃居仁). (2004). Text-based construction and comparison of domain ontology: A study based on classical poetry, *Proc. of the 18th Pacific Asia Conf. on Language,* Information and Computation, pp. 17–20.

Jiang, S.-Y. (蔣紹愚). (2003). 'Moon' and 'Wind' in Li Bai's and Du Fu's poems – Using computers for studying classical poems, *Proc. of the 1st Int'l Conf. on Literature and Information Technologies.*

Lee, J. (李思源). (2012). A classical Chinese corpus with nested part-of-speech tags, *Proc. of the 6th EACL Workshop on Language Technology for Cultural Heritage,* Social Sciences, and Humanities, pp. 75–84.

Liu, Ch.-L. (劉昭麟). (2015). Hongsu Wang, Chu-Ting Hsu, Wen-Huei Cheng, and Wei-Yun Chiu. Color aesthetics and social networks in complete Tang poems: Explorations and discoveries, *Proc. of the 29th Pacific Asia Conference on Language,* Information and Computation, pp. 132–41.

Liu, Z. (劉尊明) and Zhaopeng, W. (2012). *Quantitative Analysis of Ci in Tang and Song Dynasties.* Beijing: Peking University Press.

Lo, F. (羅鳳珠). (2005). Design and applications of systems for word segmentation and sense classification for Chinese poems, *Proc. of the 4th Conference on Technologies for Digital Archives.* (in Chinese).

Lo, F. (羅鳳珠). (2008). The research of building a semantic category system based on the language characteristic of Chinese poetry, *Proc. of the 9th Cross-Strait Symposium on Library Information Science.* (in Chinese).

Tao, W.-P. (陶文鵬). (1999). Research on Tang poems in the first half of the twentieth century, *Journal of Hubei University* (Philosophy and Social Science), v. 5. (in Chinese). http://www.guoxue.com/master/wangpijiang/wpj02.htm.

Wong, Y. T. (黃仁達). (2011). *The Colors of China.* Taipei: Linking Publishing.

Yan, R. (嚴睿). (2013). Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li, poet: Automatic Chinese poetry composition through a generative summarization framework under constrained optimization, *Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence*, pp. 2197–2203.

Zhou, M. (周明). (2009). Long Jiang, and Jing He, Generating Chinese couplets and quatrain using a statistical approach, *Proc. of the 23rd Pacific Asia Conf. on Language,* Information and Computation, pp. 43–52.

Notes

[1] This research was supported in part by the Ministry of Science and Technology of Taiwan, under the grant 104-2221-E-004-005-MY3.

[2] Bai Juyi: 白居易; Liu Changqin: 劉長卿

[3] 白髮: bai2 fa3; 青衫: qing1 shan1; 青雲: qing1 yun1; 丹砂: dan1 sha1; 青山: qing1 shan1; 白首: bai2 shou3; 紅塵: hong2 chen2; 滄洲: cang1 zhou1 ; 青春:qing1 chun1

[4] 白雲: bai yun, white cloud

[5] Collocations of "白雲" and "黃葉" (huang2 ye4, yellow leaves) appeared in poems of 劉長卿 (4 times), 盧綸 (1), 常袞 (1), and 賈島 (1).

[6] Collocations of "白雲" and "滄海" (cang1 hai3, broad ocean) appeared in poems of 劉長卿 (4), 姚合 (1), 崔峒 (1), and 賈島 (1).

[7] Collocations of "白雲" and "清露" (qing1 lou4, light dew) appeared in poems of 權德輿 (1) and 賈島 (1).

[8] Collocations of "白雲" and "流水" (liu2 shui3, running water) appeared in poems of 劉禹錫 (1), 姚合 (1), 皇甫冉 (1), 皇甫曾 (1), 賈島 (1), and 錢起 (1).

[9] Two examples by 劉長卿: "白雲留永日，黃葉減餘年" and "近北始知黃葉落，向南空見白雲多".

[10] 金: jin1, gold

[11] 白: bai1, white; 青: qing1, blue; 紅: hong2, red; 黃: huang1, yellow; 碧: bi4, green; 綠: lu4, green

[12] 紅塵: hong2 chen2; 殘紅: can2 hong2; 紅妝: hong2 zhuang1; 紅葉: hong2 ye4; 紅袖: hong2 xui4; 紅日: hong2 ri4; 紅樓: hong2 lou2

[13] 紅塵: hong2 chen2; 惺忪: xing1 song1; 空門: kong1 men2; 武皇: wu3 huang2

[14] For instance, in〈浣沙溪〉of 周邦彥 we read " 薄薄纱厨望似空。簟纹如水浸芙蓉。起来娇眼未惺忪".

[15] 酴醾: tu2 mi2; also written as "酴醿" in *QTS*.

Evaluating Modal Use in News Corpus for Constructing Rhetorical Context of Historical Event

Jyi-Shane Liu
jyishane.liu@gmail.com
National Chengchi University, Taiwan, Republic of China

Ching-Ying Lee
cylee@ukn.edu.tw
University of Kang Ning, Taiwan, Republic of China

Ke-Chih Ning
floater.xkernel@gmail.com
National Chengchi University, Taiwan, Republic of China

1. Introduction

Language use performs as a screen or filter to reality, reflecting speakers' perception and organization of the world around them (Wardhaugh, 2002). In language, modality deploys various kinds of meaning filter (types of modal expression) which variously color and modify our conceptualizations of the world and enable us to represent it with such purposeful diversity (Hoye, 2005). Modality acts as stance-taking or attitudinal qualifications, e.g. necessity (must, should) and possibility (can, may), expressing the speaker's opinion of a proposition or a predicate and its subject. Simon-Vandenbergen (1997) shows that modal certainty is an important feature of the discourse of political speakers and give a functional explanation of modal selections in political interviews. Garzone (2013) conducts a corpus-based study to show the decline in the use of "shall" in U.K. legislative texts and the use of other modal/non-modal substitutes for the somewhat offensive "shall." In this paper, we examine modality use in a corpus of historical news to observe the rhetorical stance of government propaganda at a time of governance crisis. The results indicate that a strong sense of moral persuasion and demand was manifested by significant modal use for social responsibility.

2. Theoretical framework and methodology

Theoretical studies to modality include generative, cognitive-pragmatic, and typological approaches (Hoye, 2005). The central notions to linguistic modality in typological sense are possibility and necessity (Lyons, 1995). We adopt Li's (2004) Chinese modal system which was derived from an English modality framework (van der Auwera and Plungian, 1998).

The semantic categories of the modal types and the primary Chinese modal verbs are listed below. Chinese modal verbs are poly-functional, each may indicate more than one modal senses.

- Epistemic uncertainty: estimates whether something will become a fact or not and suggests objective possibility, which corresponds to epistemic possibility. The modal verbs in Mandarin Chinese that express epistemic possibility are 能 neng2 (can), 能夠 neng2 gou4 (can), 會 hui4 (may), 可 ke3 (may), 可以 ke3 yi3 (may), 得 de2 (can).
- Epistemic probability: predicts the necessity about a finite event or state or concludes about the necessity of a current event. The modal verbs that are used to express a note of conjecture include 該 gai1 (should), 應該 ying1 gai1 (should), 要 yao4 (will), 得 dei3 (must).
- Ability: expresses subjective possibility of participant related ability, function, property, or quality. Chinese modal verbs indicating participant-internal possibility include 能 neng2 (can), 能夠 neng2 gou4 (can), 會 hui4 (can), 可 ke3 (can), 可以 ke3 yi3 (can), 得 de2 (can).
- Need: concerns with subjective necessity that corresponds to need internal to the participant involved in the state of affairs. It relates to hope, intention, and interest which come to cause the ultimate action or event. The Chinese modal verbs identified for expressing need are 要 yao4 (need), 需要 xu1 yao4 (need), 須 xu1 (need), 必須 bi4 xu1 (must), and 得 dei3 (must).
- Permission and circumstantial possibility: deals with possibility out of deontic sources like rules, regulations, authority, or non-deontic objective circumstances. Chinese modal verbs that express the notional categories of permission and circumstantial possibility are: 能 neng2 (can), 能夠 neng2 gou4 (can), 可 ke3 (may), 可以 ke3 yi3 (may), 得 de2 (can).
- Obligation and circumstantial necessity: involves deontic necessity out of morality or social conventions, and non-deontic necessity out of objective situations and reasons. Modal verbs that express obligation and circumstantial necessity include 要 yao4 (must), 該 gai1 (should), 應該 ying1 gai1 (should), 應 ying1 (should), 當 dang1 (should), 應當 ying1 dang1 (should), 須 xu1 (must), 必須 bi4 xu1 (must), 得 dei3 (must).

The corpus for our investigative purpose is the 228 event Taiwanese news archive, published by the 228 Event Memorial Foundation to compile local news articles during a short period, dated from 2/28/1947 to 5/15/1947, of widespread riot after Chinese takeover of Taiwan at the end of World War Two. The current study focuses on news articles from Taiwan Shin Sheng Daily News (TSSDN), controlled by the government at that time, while all private news publishers were shut down one week after the incident erupted. As a baseline benchmark, we use the Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus), designed to be a balanced collection from different areas of genre, style, mode, topic, and source (Huang and Chen, 1992). The two corpora are both Mandarin (modern Chinese) without notable language variants and are considered as comparable.

For both TSSDN and Sinica corpora, we use CKIP segmentation and part-of-speech (POS) tagging for first-line processing (Chang & Chen, 1995). The size of the TSSDN corpus is a total of 0.237 million word tokens. The Sinica Corpus (version 3.1), dated approximately from 1981 to 1997, contains a total of 5.738 million word tokens, which is about 24 times the size of the TSSDN corpus. For the purpose of comparing modal use in the TSSDN corpus and the Sinica corpus, the following procedure was used to extract and observe respective modal use distribution.

Step 1. For each word token in the list of common modal verbs, extract sentences that contain the specified word token with the POS tag of auxiliary verb.

Step 2. Rank the modal verbs by its occurrence frequency, i.e., the number of the extracted sentences for each modal verb.

Step 3. Exclude the bottom half of the list of modal verbs that show insignificant use and compare the use of significant modal verbs with both absolute frequency and normalized frequency (per million word tokens).

Step 4. For the TSSDN corpus, assess each sentence in use of a modal verb and determine its actual modal type.

Step 5. Compile the actual use frequency of the modal types in the TSSDN corpus.

Step 6. Compile the top five verb semantics following the use of each modal verb for observing the association of modal sense and semantic notion.

3. Results and discussion

An initial frequency observation on the list of common modal verbs in the TSSDN corpus is shown in Table 1. This led us to focus on the top half of the modal verbs that are clearly of more significant use in the investigated corpus.

Rank 1 ~ 4	Rank 5 ~ 8	Rank 9 ~ 12	Rank 13 ~ 16
應 ying1 (should) 535	可以 ke2 yi3 (can, may) 206	會 hui4 (may, can) 77	應當 ying1 dang1 (should) 21
要 yao4 (will, must) 476	須 xu1 (must) 170	得 de2, dei3 (can, must) 69	能夠 neng2 gou4 (can) 13
可 ke3 (can, may) 416	應該 ying1 gai1 (should) 138	需要 xu1 yao4 (need) 68	該 gai1 (should) 11
能 neng2 (can) 403	必須 bi4 xu1 (must) 123	當 dang1 (should) 66	需 xu1 (need) 6

Table 1. Rank list of common modal verbs by frequency in TSSDN corpus

Modal Verbs	TSSDN Corpus		Sinica Corpus		Increase Ratio	Sinica corpus Newspaper subset		Increase Ratio
	Absolute Frequency	Normalized Frequency	Absolute Frequency	Normalized Frequency		Absolute Frequency	Normalized Frequency	
應ying1 (should)	535	2257.4	3250	566.4	398.5%	1228	728.6	309.8%
要yao4 (will, must)	476	2008.4	15783	2750.7	73.0%	3135	1860.2	108.0%
可ke3 (can, may)	416	1755.3	8318	1449.7	121.1%	2439	1 447.2	121.3%
能neng2 (can)	403	1700.4	10867	1893.9	89.8%	3038	1802.6	94.3%
可以 ke2 yi3 (can, may)	206	869.2	9546	1663.7	52.2%	1981	1175.4	73.9%
須xu1 (must)	170	7 1 7.3	786	137.0	523.6%	237	140.6	510.1%
應該 ying1 gai1 (should)	138	582.3	2787	485.7	119.9%	522	309.7	188.0%
必須 bi4 xu1 (must)	123	519.0	3181	554.4	93.6%	788	467.6	111.0%

Table 2. Benchmark comparison of modal verbs use

Modal Verbs	TSSDN Corpus		Sinica corpus		Increase Ratio	Sinica corpus Newspaper subset		Increase Ratio
	Absolute Frequency	Normalized Frequency	Absolute Frequency	Normalized Frequency		Absolute Frequency	Normalized Frequency	
應ying1, 應該ying1 gai1 (should)	673	2839.7	6037	1052.1	269.9%	1750	1038.4	273.5%
可ke3, 可以 ke2 yi3 (can, may)	622	2624.5	17864	3113.3	84.3%	4420	2622.6	100.1%
要yao4 (will, must)	476	2008.4	15783	2750.7	73.0%	3135	1860.2	108.0%
能neng2 (can)	403	1700.4	10867	1893.9	89.8%	3038	1802.6	94.3%
須xu1, 必須bi4 xu1 (must)	293	1236.3	3967	691.4	178.8%	1025	608.2	203.3%

Table 3. Benchmark comparison of modal verbs use in semantic notions

Modal Type by Modal Verb	Epistemic uncertainty	Epistemic probabilisty	Ability	Need	Circumstantial possibility	Circumstantial Reed	Permission	Obligation
應ying1, 應該ying1 gai1 (should)		11				15		647
可ke3, 可以ke2 yi3 (can, may)	112		97		248		165	
要yao4 (will, must)		27		74		33		342
能neng2 (can)	8		168		138			
須xu1, 必須bi4 xu1 (must)				3		77		213
Absolute Frequency	120	38	265	77	386	125	254	1202
Normalized Frequency	506.3	160.3	1118.1	324.9	1628.7	527.4	1071.7	5071.7

Table 4. Frequency distribution of modal type expression by modal verbs

Modal Verb	Top Five Verb Semantics with Occurrence Frequency and Ratio	Semantic Meaning
應ying1, 應該ying1 gai1 (should)	注意zhu4 yi4 (33) (4.9%) 遵守zun1 shou3, 遵照 zun1 zhao4 (27) (4.0%) 負責fu4 ze2 (27) (4.0%) 處分chu3 fen4, 嚴懲yan2 cheng3 (27) (4.0%) 檢查 jian3 cha2, 調查diao4 cha2 (25) (3.7%)	Reed Comply responsible punish inspect
可ke3, 可以ke2 yi3 (can, may)	說shuo1, 知zhi 1(63) (10.1%)恢復hui1 fu4, 穩定 wen3 ding4 (45) (7.2%) 實現shi2 xian4, 達成da2 cheng2 (36) (5.8%) 運輸yun4 shu1, 通行tong1 xing2 (32) (5.1%) 報告bao4 gao4, 提請ti2 qing3 (26) (4.2%)	speak, know restore, stabilize achieve transport, pass report, submit
要yao4(will, must)	知道(zhi1 dao4), 認識(ren4 shi4) (64) (13.4%) 努力nu3 li4, 加強jia1 qiang2 (34) (7.1%) 說明shuo1 ming2 (21) (4.4%) 檢舉jian3 ju3, 糾正jiu1 zheng4 (17) (3.6%) 負責fu4 ze2 (14) (2.9%)	know, perceive strive, strengthen Explain report fault, correct responsible
能neng2 (can)	了解le3 jie3, 明瞭ming2 liao3 (39) (9.7%) 恢復hui1 fu4 (24) (6.0%) 實現shi2 xian4, 達成da2 cheng2 (22) (5.4%) 解決jie3 jue2, 克服ke4 fu2 (19) (4.7%) 看懂kan4 dong3, 讀寫du2 xie3 (16) (4.0%)	under stand Restore achieve solve, overcome read/write
須xu1, 必須bi4 xu1 (must)	登記deng1 ji4, 註冊zhu4 ce4 (12) (4.1%) 持有chi2 you3 (12) (4.1%) 注意zhu4 yi1 (11) (3.8%) 懲辦 cheng2 ban4, 處分chu3 fen4 (8) (2.7%) 肅清su4 qing1, 鎮壓zhen4 ya1 (7) (2.4%)	register carry (valid permit) Reed punish exterminate, suppress

Table 5. Primary verb semantics following modal verbs

Table 2 compares the use of major modal verbs in TSSDN corpus, Sinica corpus, and the newspaper subset of Sinica corpus (29.4% of the whole corpus). It is observed that TSSDN corpus shows a considerable frequency variation of modal verb use when compared with Sinica corpus and its newspaper subset. Among the eight modal verbs, three pairs of modal verbs are actually variants of each other and may be aggregated to better represent the semantic notions of the modal verbs.

In Table 3, we aggregate three pairs of modal verb variants of the same semantic notion and re-calibrate the relative amount of modal use in the two corpora and one sub-corpus. The comparison shows that the use of 應 ying1, 應該 ying1 gai1 (should) and 須 xu1, 必須 bi4 xu1 (must) in TSSDN corpus are significantly frequent than in Sinica corpus and its newspaper sub-corpus, while the use of the other modal verbs are somewhat comparable. This indicates that TSSDN corpus contains a strong attitude and stance through the unusual emphasis of should and must.

Next, we observe how the use of modal verbs is distributed in the modal system to depict various aspects of attitude and stance. Each occurrence of a modal verb in a sentence is categorized in modal type by independent coders. Disputed codes are discussed to reach consensus decision. Table 4 breaks down the occurrence of modal type expression by the poly-functional modal verbs. The results reveal an extremely high concentration on the modality type of obligation, signaling a heavy dose of demand and persuasion of social responsibility from government propaganda.

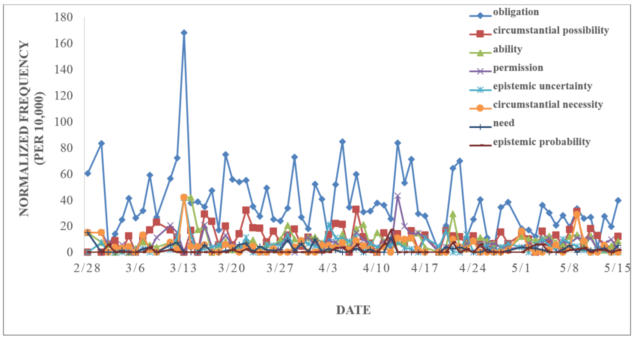


Figure 1. Chronological occurrence of modal types on a daily timeline

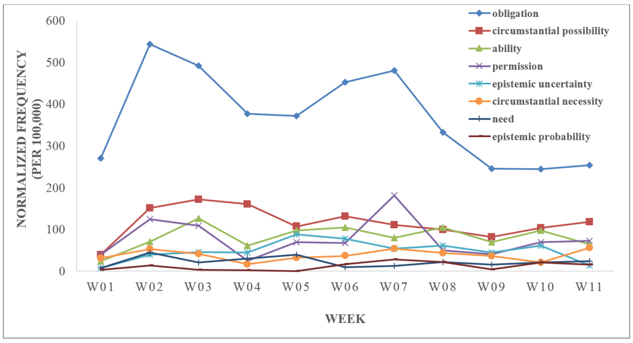


Figure 2. Chronological occurrence of modal types on a weekly timeline

We also observe the normalized occurrence with respect to word count of news reports on a daily timeline in Figure 1 and weekly timeline in Figure 2. The temporal variation depicts a process of employing the rhetoric of obligation that immediately peaks in the second week, followed by a lower peak in the seventh week before gradually reduced in the third month, over the period in which social order was lost and regained, social activities was disrupted and restored.

Table 5 compiles the top five verb semantics, ranked by the occurrence frequency and ratio, associated with the use of each modal verb in a sentence. This helps provide a better rhetorical picture of what is being said, appealed, urged, or even warned. Overall, we observe a rhetoric sense of strict attitude and firm stance on exercising and restoring control of social order.

In conclusion, our study seems to indicate that modality is an effective linguistic feature for extracting narrative stance and provides a convenient contextual view of a corpus. Our future work includes examining more comprehensive modal expression and evaluating against corpora of various historical context.

Bibliography

Chang, L. P. and Chen, K. J. (1995). The CKIP part-of-speech tagging system for modern Chinese texts. *Proceedings of ICCPOL'95*. Hawaii, U.S.A.

Garzone, G. (2013). Variation in the use of modality in legislative texts: Focus on shall. *Journal of Pragmatics*, **57**: 68-81.

Hoye, L. F. (2005). "You may think that; I couldn't possibly comment!" Modality studies: Contemporary research and future directions. Part I. *Journal of pragmatics*, **37**(8): 1295-321.

Huang, C. R. and Chen, K. J. (1992). A Chinese corpus for linguistics research In *Proceedings of the 1992 International Conference on Computational Linguistics* (COLING-92). Nantes, France, pp. 1214-17.

Li, R. Z. (2004). *Modality in English and Chinese: A Typological Perspective*. Doctoral Dissertation. University of Antwerp.

Lyons, J. (1995). *Linguistic Semantics: An Introduction*. Cambridge University Press.

Simon-Vandenbergen, A. M. (1997). Modal (un)certainty in political discourse: A functional account. *Language Sciences*, **19**(4): 341-56.

Van der Auwera, J. and Plungian, V. A. (1998). Modality's semantic map. *Linguistic Typology*, **2**: 79-124.

Wardhaugh, R. (2002). *An Introduction to Sociolinguistics*. Blackwell Publishing.

CORE - A Contextual Reader based on Linked Data

Eetu Mäkelä
eetu.makela@aalto.fi
Aalto University, Finland

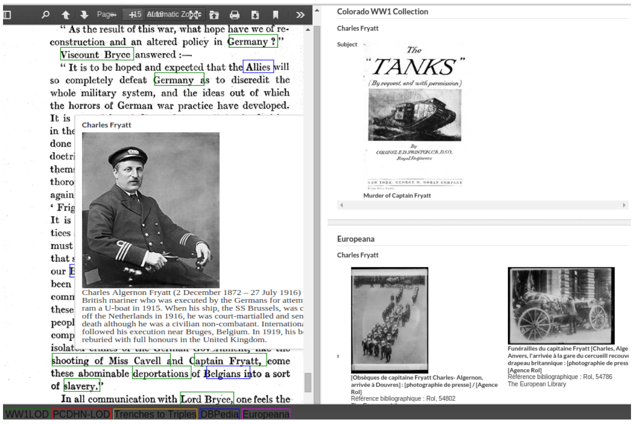
Thea Lindquist
thea.lindquist@colorado.edu
University of Colorado Boulder, United States

Eero Hyvönen
eero.hyvonen@aalto.fi
Aalto University, Finland

Motivation

In a relatively recent study on the needs of humanities faculty and students in using digital sources (Lindquist and Long 2011), two major issues were identified: 1) locating data relevant to a topic when online collections are distributed across institutions and systems; and 2) being able to explore the items found in context. In addition, problems were identified with crossing language barriers and with ambiguities and variants in names.

The CORE contextual reader is an application that uses natural language processing and Linked Data (Heath and Bizer 2011) techniques to address these issues in the context of close reading of primary source material[1]. Particularly, the CORE application has been designed to improve the user reading experience with texts in a domain not entirely familiar to them. Examples of this situation include a history student approaching a new topic through primary sources, or a layperson trying to make sense of law texts.



" As the result of this war, what hope have we of reconstruction and an altered policy in Germany ? "
Viscount Bryce answered :—
" It is to be hoped and expected that the Allies will so completely defeat Germany as to discredit the whole military system, and the ideas out of which the horrors of German war practice have developed. It is
in the
done
doctri
thems
thoro
again
' Frig
It is
tices
must
that s
our E
been
comn
these
peopl
comp
isolated
shooting of Miss Cavell and Captain Fryatt, come these abominable deportations of Belgians into a sort of slavery.'
In all communication with Lord Bryce, one feels the

Charles Fryatt

Charles Algernon Fryatt (2 December 1872 – 27 July 1916)
British mariner who was executed by the Germans for attem
ram a U-boat in 1915. When his ship, the SS Brussels, was c
off the Netherlands in 1916, he was court-martialled and sen
death although he was a civilian non-combatant. Internationa
followed his execution near Bruges, Belgium. In 1919, his b
reburied with full honours in the United Kingdom.

Murder of Captain Fryatt

[Obsèques de capitaine Fryatt Charles- Algernon, arrivée à Douvres] : [photographie de presse] / [Agence Rol]
Référence bibliographique : Rol, 54802

Funérailles du capitaine Fryatt [Charles, Alge
Anvers, l'arrivée à la gare du cercueill recouv
drapeau britannique : [photographie de press
[Agence Rol]
Référence bibliographique : Rol, 54786
The European Library

Figure 1. The contextual reader interface

The CORE user interface

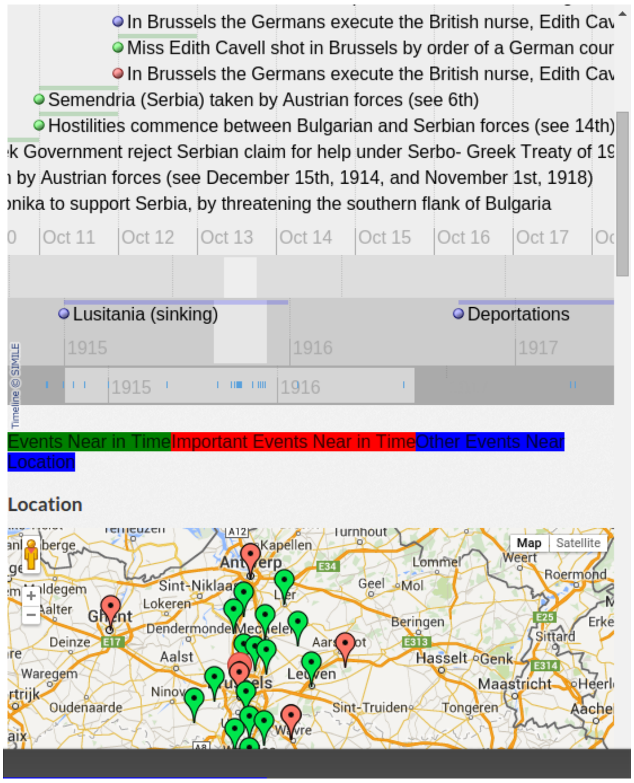
CORE supports contextualization in and understanding of unfamiliar documents by utilizing Linked Data reference vocabularies and datasets to identify entities in any PDF file or web page. For each discovered entity, CORE can then present configurable information sourced from these reference datasets on a mouse-over inside the web browser being used to read the document. Figure 1 shows this functionality in the context of reading a primary source document dealing with the First World War. The document, a scanned PDF, is shown in the interface on the left-hand side. Colored boxes highlight all of the entities identified by CORE. Here, the user has moused over "Captain Fryatt", and the interface has brought up his picture and a short biography. Other examples of contextual information shown are word definitions for domain-specific vocabulary, maps showing the geographical context of unfamiliar places mentioned, and so on.

If further information is needed, an entity can be clicked on to load more information and context into the pane on the right-hand side of the reader. In this pane, contextualization is further supported by visualizations, for example, locating the entity of interest temporally on a timeline and geographically on a map. Figure 2 shows these visualizations for an identified event, in this case the execution of Nurse Edith Cavell by the Germans in Belgium during WW1. At the top of the pane, the event is contextualized temporally among other war events. These are color-coded to differentiate: 1) important top-level wartime events sourced from the Imperial War Museum, 2) all events happening in the same timeframe, and 3) other wartime events happening nearby. Below the timeline, all of these events are presented on a map to give a geographical perspective. Clicking on any of the entities visualized loads the information pertaining to that entity into the contextualization pane, allowing further navigation of the context.

In addition to providing more nuanced context, the right-hand pane of CORE also facilitates serendipitous discovery of further related content. Using the configured Linked Data vocabularies, CORE is able to extract relevant search terms for an entity of interest. These search terms can then be used to discover related content from configured endpoints, even if they support only simple text searches. In Figure 1, this functionality is seen on the right-hand side of the user interface. First, formally encoded metadata brings in another relevant primary source from the University of Colorado Boulder's (CU-Boulder) WWI Collection Online[2]. Images of the burial of Captain Fryatt from Europeana[3], on the other hand, are found not through formally encoded keywords, but rather a match on his name that appears in the textual description of the images.

Among the extracted terms used in the query are multilingual labels for places, variant names for actors and events, etc. Leveraging these terms enables CORE to cross language barriers and handle naming variations. To improve recall even further, the search term extraction can be configured to include terms for related entities, such as the actors participating in an event or the names of all villages in a particular municipality under investigation.

Because CORE is able to dynamically process most HTML and PDF content, any linked resource can be loaded into the contextual reader by clicking on it. This function facilitates endless browsing on a topic through thematic and contextual connections, regardless of from where the linked material comes.



● In Brussels the Germans execute the British nurse, Edith Cav
● Miss Edith Cavell shot in Brussels by order of a German cour
● In Brussels the Germans execute the British nurse, Edith Cav
● Semendria (Serbia) taken by Austrian forces (see 6th)
● Hostilities commence between Bulgarian and Serbian forces (see 14th)
k Government reject Serbian claim for help under Serbo- Greek Treaty of 19
ı by Austrian forces (see December 15th, 1914, and November 1st, 1918)
ɔnika to support Serbia, by threatening the southern flank of Bulgaria

🟩 Events Near in Time 🟥 Important Events Near in Time ▇ Other Events Near Location

Location

Figure 2. Contextual visualizations for the shooting of Nurse Edith Cavell

System demonstrators

In contrast to most other similar systems, a CORE instance can relatively easily (and always should be) configured for a particular domain, thus ensuring the contextual information provided is actually useful and interesting to the end-user.

To provide its services, CORE makes use of dynamic, configurable entity recognition, in which modular lexical analysis services are combined with SPARQL queries[4]. This allows multilingual entity recognition against any vocabulary stored at a Linked Data endpoint. A configuration therefore consists of tuning the lexical analysis service to a particular domain and set of languages, as well as defining the endpoints and queries to be used in bringing in contextual information and related resources.

While in future the application is intended to be fully configurable using a web user interface, currently new instances must be configured from the source code, released under the MIT open source license at http://github.com/jiemakel/core/. Thus far, three different demonstrators have been created[5].

The first of these is the contextual reader for First World War primary sources available at http://demo.seco.tkk.fi/ww1/. For vocabularies, it draws on the WW1LOD dataset (Mäkelä et al. 2015), the vocabularies of 1914–1918 Online[6], the Europeana 1914–1918[7] thesaurus, the Out of the Trenches (Pan-Canadian Documentary Heritage Network 2012) and Trenches to Triples[8] vocabularies, and DBpedia (Lehmann et al. 2015). Repositories used for sourcing related content are CU-Boulder's WWI Collection Online, WW1 Discovery[9], Europeana, the Digital Public Library of America (DPLA)[10], and The European Library (TEL)[11].

To further demonstrate multilingual support as well as support for inflected languages, a second contextual reader has been configured to support the study of ancient Roman sources, be they translated into English or still in the original Latin. This installation is available at http://demo.seco.tkk.fi/ancore/. Here, ancient place names are located on maps through the Pleiades gazetteer of ancient places[12], while information on entities like people and mythical characters mentioned in the texts is sourced from the English and Latin DBpedias. Targeted repositories are the Perseus Catalog[13], the various Pelagios datasets[14], and again Europeana, DPLA, and TEL.

The final demonstrator, aimed at supporting the reading of legal documents in the highly inflected language of Finnish, is available at http://demo.seco.tkk.fi/laki/. In this case, the documents are drawn from, for example, the consolidated legislation[15] and the precedents of Finnish supreme courts[16] published by the Finnish Ministry of Justice. In addition to linking these distributed resources to one another, the application is able to bring in news articles[17] dealing with laws of interest published by Edita Publishing.

When reading documents containing precise legal terminology, the reader is supported by definitions from the legal terminology section of the Bank of Finnish Terminology in Arts and Sciences[18], the Asseri vocabulary of the Ministry of Justice, the Edilex legal vocabulary from Edita, the Finnish law vocabulary from Talentum Publishing, and the legal terminology section of the Finnish DBpedia. In addition to Finnish, this reader has also been configured with limited support for Swedish, as Finland is a bilingual country.

Conclusions and future work

The CORE contextual reader clearly demonstrates the potential of utilizing Linked Data vocabularies to bridge institutional silos and language barriers, even in situations where the structured metadata of the corresponding databases is lacking. On the other hand, the core mission of the tool is to support contextualization and understanding.

While initial experience points both to significant overall support, as well as a marked increase in support with regard to less domain-configured alternatives (Csomai and Mihalcea 2007; Olango, Kramer, and Bouma 2009), a formal user evaluation of the reader remains to be conducted. This will be the natural next step for the project, and plans for testing the WW1 version of the reader are already underway.

At the same time, the CORE reader is currently seeing uptake in new contexts, most notably a project to unify disparate material related to the Finnish view of the Second World War (Hyvönen et al. 2015). Supporting these new contexts may require further development of components of the reader. For example, the Second World War material under study contains multiple distinct places and people with the same names. To properly handle these would require better support for disambiguation in the entity recognition component of the reader.

Bibliography

Csomai, A. and Mihalcea, R. (2007). Linking Educational Materials to Encyclopedic Knowledge. *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*. Amsterdam: IOS Press, pp. 557–59.

Dzbor, M., Motta, E. and Domingue, J. (2007). Magpie: Experiences in supporting Semantic Web browsing. *Web Semantics: Science, Services and Agents on the World Wide Web.*

Heath, T. and Bizer, Ch. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

Hyvönen, E. et al. (2015). Second World War on the Semantic Web: The WarSampo Project and Semantic Portal. *Proceedings of 14th International Semantic Web Conference 2015 (ISWC 2015), Posters and Demos*. Forthcoming. Bethlehem, PA, USA: CEUR-WS Proceedings.

Lehmann, J. et al. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, **6**(2): 167–95.

Lindquist, T. and Long, H. (2011). How can educational technology facilitate student engagement with online primary sources?: A user needs assessment. *Library Hi Tech*, **29**(2): 224–41.

Mäkelä, E. (2014). Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text. *Proceedings of the ESWC 2014 demonstration track, Springer-Verlag.*

Mäkelä, E. et al. (2015). World War 1 as Linked Open Data. Submitted for review.

Olango, P., Kramer, G. and Bouma, G. (2009). TermPedia for interactive document enrichment using technical terms to provide relevant contextual information. *International Multiconference on Computer Science and Information Technology, 2009. IMCSIT '09*, pp. 265–72.

Pan-Canadian Documentary Heritage Network (2012). *PCDHN Linked Open Data Visualization "Proof-of-Concept"*. Tech. rep.

Notes

[1] In this, the application can be seen as a spiritual successor to the Magpie tool (Dzbor, Motta, and Domingue 2007), although the two share nothing concrete with each other.

[2] http://cudl.colorado.edu/luna/servlet/UCBOULDER-CB1~58~58

[3] http://europeana.eu/

[4] For technical details of the system, see (Mäkelä 2014).

[5] For those technically oriented, the configuration files for these demonstrators can be perused at https://github.com/jiemakel/core/blob/v1.0.0/app/scripts/main/.

[6] http://www.1914-_1918-_online.net/

[7] http://www.europeana1914-_1918.eu/

[8] http://data.aim25.ac.uk/about_t3.php

[9] http://ww1.discovery.ac.uk/

[10] http://dp.la/

[11] http://www.theeuropeanlibrary.org/

[12] http://pleiades.stoa.org/

[13] http://catalog.perseus.org/

[14] http://pelagios.dme.ait.ac.at/api/datasets

[15] http://finlex.fi/fi/laki/ajantasa/

[16] http://finlex.fi/fi/oikeus/

[17] http://www.edilex.fi/uutiset

[18] http://tieteentermipankki.fi/wiki/Oikeustiede

Khepri - a Modular View-Based Tool for Exploring (Historical Sociolinguistic) Data

Eetu Mäkelä
eetu.makela@aalto.fi
Aalto University, Finland

Tanja Säily
tanja.saily@helsinki.fi
University of Helsinki, Finland

Terttu Nevalainen
terttu.nevalainen@helsinki.fi
University of Helsinki, Finland

Motivation

Digital humanities needs tools that better support the core processes of humanistic inquiry. This includes support for handling uncertainty and incompleteness in the data, for interactive exploration, and for fluidly moving between close and distant reading (Drucker 2011; Jänicke et al., 2015; Caviglia, Ciuccarelli, and Coleman, 2012; Uboldi and Caviglia, 2015).

The Khepri tool presented here is part of a project to develop a modular set of components that take these requirements into account, and can be connected and configured to respond to the needs of a particular humanities task and data. Khepri targets data stored as Linked Data (Heath and Bizer, 2011), a set of scalable standards that has gained widespread adoption particularly in the sphere of cultural heritage.

Development process

To ensure the tools developed meet the needs of humanities users, they are being developed iteratively, utilizing participatory design in case studies, as advocated by the field of design science (Hevner et al., 2004; Peffers et al., 2007; Wieringa 2009). The task of the computer scientist is to see beyond these individual studies; to identify common components allowing the tools to generalize beyond the projects under scrutiny.

To date, a variety of collaborations have been embarked upon, from the prosopographical study of the Republic of Letters[1], through supporting engagement with WW1 primary sources (Mäkelä, Törnroos, et al., 2015), to developing a contextual network for Finnish fiction (Mäkelä, Hypén, and Hyvönen, 2013). Together, these span a range of research questions, and types of data.

Through these collaborations, a prevalent process of inquiry was identified – the need to explore and contrast differently constrained subsets of a dataset. For instance, this might be looking at the correspondence networks of different individuals and comparing them, or looking at how possible values of a linguistic variable behave with respect to each other as well as associated metadata.

To support this process, Khepri utilizes the view-based paradigm (Mäkelä, 2010), where data is presented simultaneously from different perspectives, with each perspective acting both as a visualization as well as a means to constrain what is shown. A proper implementation of the paradigm also allows for speedy informed variation of parameters, and thus interactive exploration.

Because the views interact in a defined way, they can be developed as separate components targeting major visualization classes such as geographical, temporal or statistical. Each individual Khepri instance can then select from these the views suitable for that particular use.

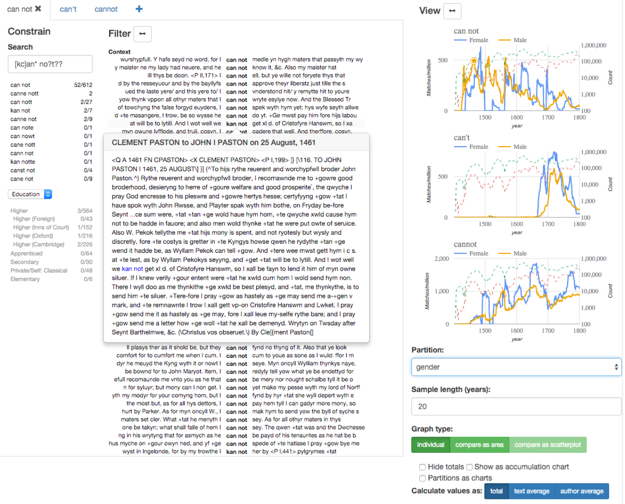
Thus far, most of the work has been preparatory, with the functionalities simulated through ad-hoc disconnected components, tied together and supplemented by manual work of the computer scientist. However, now a first complete tool for a particular task has been developed. This instance has been configured for historical sociolinguistics.

Khepri for historical sociolinguistics

Historical sociolinguistics is the study of language in relation to social factors through time (Nevalainen and Raumolin-Brunberg, 2003). A possible research question would be to chart the role of gender, age and socioeconomic status in the diffusion of the English progressive (as in *I am writing*). From the viewpoint of the Khepri tool, this is interesting because it requires combining access to unstructured text with access to the structured (meta) data describing their authors.

This is also the area where current tools fall short, for while corpus tools (e.g. CQPweb (Hardie, 2012), Korp (Borin, Forsberg, and Roxendal, 2012) and WordSmith[2]) enable querying texts by linguistic features, they poorly support walking from the texts to the attributes of the authors. On the other hand, tools for visually exploring structured data (e.g. Palladio[3], Europeana4D[4] and RAW[5]) do not support interacting with text corpora.

This makes research currently very labor-intensive. For instance, if one wishes to study the aforementioned progressive, one first searches for instances of *-ing* in the corpus using a corpus tool. The instances are then exported into Excel to analyze them and eliminate false hits such as gerunds (*My favourite hobby iswriting*). Next, the number of hits produced by each person is calculated using another sheet that lists the authors by gender, age, socioeconomic status and time period. These numbers are then exported for statistical analysis and visualization. Because the corpus texts, spreadsheets, visualizations and statistical analyses are not connected to each other, the exploration and interpretation of the observations is cumbersome and time-consuming at every stage.



Search

[kc]an* no?t??

can not	52/612
canne nott	2
can nott	2/27
kan not	2/7
canne not	2/9
can note	0/1
can nowt	0/1
cane nott	0/1
cann not	0/1
kan notte	0/1
canst not	0/4
cane not	0/9

Education

Higher	3/564
Higher (Foreign)	0/43
Higher (Inns of Court)	1/152
Higher (Oxford)	1/216
Higher (Cambridge)	2/226
Apprenticed	0/64
Secondary	0/30
Private/Self: Classical	0/48
Elementary	0/6

Context

```
         wurshypfull. Y hafe seyd no word, for I  can not  medle yn hygh maters that passyth my wy
        y maister ne my lady had neuere, and he  can not  know it, &c. Also my maister hat
                  ill thys be doon. <P II,171> I  can not  ell, but ye wille not foryete thys that
       d by the resseyuour and by the bayllyfs  can not  approve theyr liberatz just tille the s
          ued the laste yere/ and this yere to/ I  can not  vnderstond hit/ y remytte hit to youre
        yow thynk vppon all othyr maters that I  can not  wryte esylye now. And the Blessed Tr
       of towchyng the false forgyd euydens, I  can not  spek wyth hym yet; hys wyfe seyth allwe
       d +te masangere, I trow, be so wysse he  can not  do yt. +Ge mwst pay him fore hijs labou
              at will be to lytill. And I wot well we  kan not  get xl d. of Cristofyre Hanswm, so I xa
          myn owune lyfflode, and truli, cosyn, I  can not  gadere that well. And therffore, cosyn,
```

CLEMENT PASTON to JOHN I PASTON on 25 August, 1461

<Q A 1461 FN CPASTON> <X CLEMENT PASTON> <P I,199> [} [\116. TO JOHN PASTON I 1461, 25 AUGUST\] }] (^To hijs rythe reuerent and worchypfwll broder John Paston.^) Rythe reuerent and worchypfwll broder, I recomawnde me to +gowre good broderhood, desieryng to herre of +goure welfare and good prosperite`, the qwyche I pray God encresse to his pleswre and +gowre hertys hesse; certyfyyng +gow +tat I haue spok wyth John Rwsse, and Playter spak wyth him bothe, on Fryday be-fore Seynt ...ce sum were, +tat +tan +ge wold haue hym hom, +te qwyche xwld cause hym not to be hadde in fauore; and also men wold thynke +tat he were put owte of seruice. Also W. Pekok tellythe me +tat hijs mony is spent, and not ryotesly but wysly and discretly, fore +te costys is gretter in +te Kyngys howse qwen he rydythe +tan +ge wend it hadde be, as Wyllam Pekok can tell +gow. And +tere wee mwst gett hym i c s. at +te lest, as by Wyllam Pekokys seyyng, and +get +tat will be to lytill. And I wot well we kan not get xl d. of Cristofyre Hanswm, so I xall be fayn to lend it him of myn owne siluer. If I knew verily +gour entent were +tat he xwld cum hom I wold send hym non. There I wyll doo as me thynkithe +ge xwld be best plesyd, and +tat, me thynkythe, is to send him +te siluer. +Tere-fore I pray +gow as hastely as +ge may send me a-+gen v mark, and +te remnawnte I trow I xall gett vp-on Cristofire Hanswm and Lwket. I pray +gow send me it as hastely as +ge may, fore I xall leue my-selfe rythe bare; and I pray +gow send me a letter how +ge woll +tat he xall be demenyd. Wrytyn on Twsday after Seynt Barthelmwe, &c. (\Christus vos obseruet.\) By Cle[{ment Paston{]

```
          ll plasys ther as it shold be, but they  can not  fynd no thyng of it. Also that ye look
         comfort for to cumfort me when I cum. I  can not  cum to youe as sone as I wuld: ffor I m
         dyr he meuyd the Kyng wyth it or nowt I  can not  seye. Myn oncyll Wylliam thynkys naye,
              be bownd for to John Maryot. Item, I  can not  redyly tell yow what ye be endettyd for
       efull recomaunde me vnto you as he that  can not  be mery nor nought schalbe tyll it be o
            n for syluyr; but mony can I non get. I  can not  yet make my pesse wyth my lord of Norff
       yth my modyr for your comyng hom, but I  can not  fynd by hyr +tat she wyll depert wyth e
             the most but, as for all hys dettors, I  can not  pay hem tyll I can gadyr more mony, so
            hurt by Parker. As for myn oncyll W., I  can not  mak hym to send yow the byll of syche s
            maters set cler. What +tat he menyth I  can not  sey. As for all othyr maters in thys
            one be takyn; what shall falle of hem I  can not  sey. The qwen +tat was and the Dwchesse
           ng in his wrytyng that for asmych as he  can not  be payd of his tenauntes as he hat be b
      hus myche on +gour owyn hed, and yf +ge  can not  spede of +te hatlase I pray +gow bye me
          wyst in Ingelonde, for by my trowthe I  kan not  her by <P I,441> pylgrymes +tat
```

Partition: gender

Sample length (years): 20

Graph type: individual | compare as area | compare as scatterplot

□ Hide totals □ Show as accumulation chart

□ Partitions as charts

Calculate values as: total | text average | author average

Figure 1. the Khepri for historical sociolinguistics interface

The user interface configuration of Khepri for historical sociolinguistics

The Khepri interface for historical sociolinguistics is depicted in Figure 1. The interface is divided into three columns, with the views contained in each having different primary purposes.

On the left are views aimed primarily at producing a subset of interest. The first view is for text search. Below the query, matching keywords from the data are presented for evaluation. Notice that two sets of counts are given. One shows the overall amount of hits for a keyword in the corpus, while the other takes into account constraints set in other windows. This way, the view acts not only as a selector, but also as a statistical breakdown of the current subset.

Below the keyword search view, the user can add metadata views. Here for example, a view visualizes and allows one to constrain the data through the lens of the author's education.

The second column shows the items in the current subset. Matches are shown in their textual context, with metadata and additional context available on mouse-over. While tuned for close reading, this view also acts as a filter. Clicking on an item removes it from the current subset. For linguistic research, this is important as the inclusion or exclusion of a particular example of a phenomenon may depend on contextual cues and background knowledge that cannot be defined as search parameters, but require manual evaluation.

When focusing on close reading, the column can be expanded to occupy the whole right-hand side of the interface. Expanded, the view shows additional metadata, such as the author and year of the texts. The view can also be sorted according to these properties, as well as grouped by them, so that for example only a listing of the authors, or the linguistic types (e.g. different words ending with *-ing*) is shown, with the individual matches revealed by expanding.

To further help in keeping a close reading task organized, the interaction between this view and the constraining views has been designed so that it is easy to temporarily restrict the matches shown to only those from e.g. a particular spelling, or a particular social class.

Finally, the column on the right is intended primarily for visualization. In fact, it can visualize and contrast multiple subsets of the data. To facilitate this, the first two columns are subsumed in a tabbing container, with each tab containing the query state of a single subset. In the example of Figure 1, these are spelling variants of the negated auxiliary verb *cannot* (written separately, contraction, written together).

By default, the frequency of each subset is visualized as its own line chart. However, numerous options affecting this are provided, drawn from best practices in the field (Hinneburg et al., 2007). For example, separate lines can be graphed for each of the values of a particular metadata property. In Figure 1 for example, each chart contains lines for male and female writers, showing that the use of the form “can not” seems to follow an approximately linear decline for men, but not for women.

To prevent misinterpretations arising from small samples, each graph can be accompanied by a dotted logarithm representing the size of the corpus as a whole for that metadata value. The interface also supports bootstrapping to visualize confidence intervals. As this takes considerable time to calculate, it should only be enabled when a seemingly significant discovery needs verification.

The interface also offers alternative charts. For example, when comparing possible values of a single linguistic variable, the area chart visualization shown in Figure 2 is appropriate. In addition, a motion chart visualization (Figure 3, inspired by the static scatterplots in Nevalainen, Raumolin-Brunberg, and Mannila (2011)) is provided, used to see how different individuals relate to the variable under study, and even how they change their use through time.

In line with the view-based querying paradigm, all visualizations also act as selectors, enabling delving deeper into interesting phenomena. Through them, one can for example constrain the instance list to show only usage by women in a particular timespan, or in the case of the motion chart, even the use of a single individual.

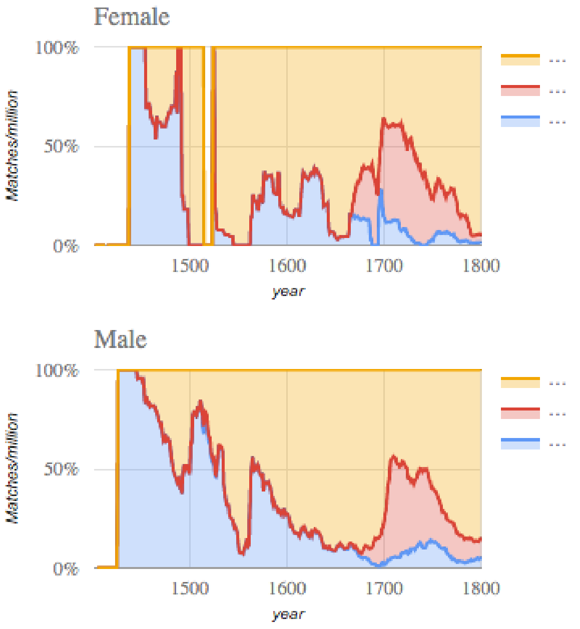


Figure 2. Area charts showing the relative proportions of “can not” (blue), “cannot” (yellow) and “can’t” (red) by time and gender

Discussion and future work

Khepri for historical sociolinguistics is the first complete version of the tool. It is also only in its second iteration, and will continue to improve based on feedback. However, it has already been received with excitement, enabling research that was previously too time-consuming to attempt.

With the architecture of the tool now in place, other instances will soon follow, targeting next the Republic of Letters and Finnish fiction use cases. This can be said because all the views created are actually generic, and can be pointed to different data by reconfiguring. For example, text search is also useful for locating individuals or books, while the metadata facets directly target structured data already. The views requiring most modification are the statistical charts, but even here work will be fine-tuning to match differing metrics. Correspondingly, any visualizations developed for other scenarios can be imported here, to for example visualize language phenomena on maps.

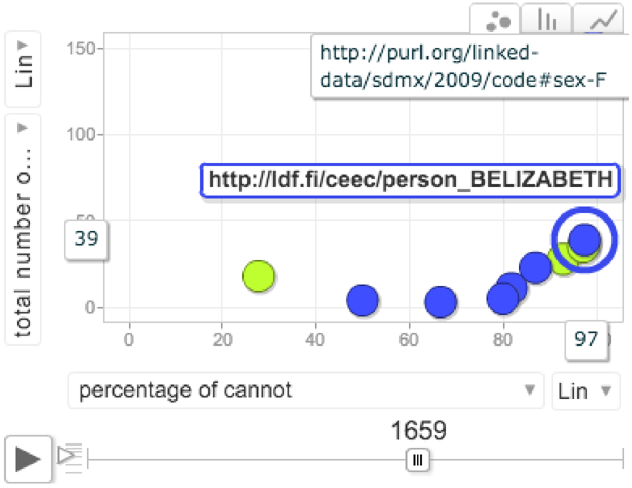


Figure 3. Motion chart showing how many percent of individual writers use the form “cannot”

Bibliography

Lars, B., Forsberg, M. and Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12).*

url: http://www.lrec-_conf.org/proceedings/lrec2012/pdf/248_Paper.pdf.

Caviglia, G., Ciuccarelli, P. and Coleman, N. (2012). Communication Design and the Digital Humanities. *Proceedings of the 4th International Forum of Design as a Process.*

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, **5**(1): 1–21.

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, **17**(3): 380–409. doi: 10.1075/ijcl.17.3.04har.

Heath, T. and Bizer, Ch. (2011). *Linked Data: Evolving the Web into a Global Data Space.*

Synthesis Lectures on the Semantic Web. Morgan and Claypool Publishers. doi: 10.2200/S00334ED1V01Y201102WBE001.

Hevner, Alan R., et al. (2004). Design Science in Information Systems Research. *MIS Quarterly*, **28**(1): 75–105.

Hinneburg, A., et al. (2007). How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change. *Literary and Linguistic Computing*, **22**(2): 137–150. doi: 10.1093/llc/fqm006.

Jänicke, S., et al. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *Eurographics Conference on Visualization (EuroVis) - STARs.*

Ed. by R. Borgo, F. Ganovelli, and I. Viola. The Eurographics Association. doi: 10.2312/eurovisstar.20151113.

Mäkelä, E. (2010). View-Based User Interfaces for the Semantic Web. D.Sc. dissertation. PhD thesis. Aalto University, School of Science and Technology, Espoo.

Mäkelä, E., Hypén, K. and Hyvönen E. (2013). *Fiction Literature as Linked Open Data - the BookSampo Dataset.*

Mäkelä, E., Törnroos, J., et al. (2015). *World War 1 as Linked Open Data*. Submitted for review.

Nevalainen, T. and Raumolin-Brunberg, H. (2003). *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Pearson Education.

Nevalainen, T., Raumolin-Brunberg, H. and Mannila H. (2011). The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change*, **23**(1): 1–43. doi: 10.1017/S0954394510000207.

Peffers, K., et al. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, **24**(3): 45–77.

Uboldi, G. and Giorgio C. (2015). Information Visualizations and Interfaces in the Humanities. English. *New Challenges for Data Design.*

Ed. by David Bihanic. Springer London, pp. 207–18. isbn: 978-1-4471-6595-8. doi: 10.1007/978-1-4471-6596-5˙11.

Wieringa, R. (2009). Design science as nested problem solving. *Proceedings of the 4th international conference on design science research in information systems and technology.* ACM, p. 8.

Notes

[1] http://www.republicofletters.net/

[2] http://www.lexically.net/wordsmith/

[3] http://palladio.designhumanities.org/

[4] http://www.tinyurl.com/e4d-_project

[5] http://raw.densitydesign.org/

Where Close and Distant Readings Meet: Text Clustering Methods in Literary Analysis of Weblog Genres

Maciej Maryl
maciej.maryl@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences, Poland

Maciej Piasecki
maciej.piasecki@pwr.edu.pl
Wrocław University of Technology

Ksenia Młynarczyk
ksenia.mlynarczyk@gmail.com
Wrocław University of Technology

Problem: towards a non-topical classification of weblog genres

The existing typologies of weblog genres - both popular and academic - are based on the blog topic, e.g. cooking blogs, travels, business (cf. Morrison 2008) or its medium, e.g. vlogs, picture logs (cf. Herring et al. 2005). In order to go beyond topical distinctions, Maryl, Niewiadomski and Kidawa (2016) conducted an interpretive study on the sample of 322 popular Polish blogs. They adopted a new-rhetorical approach, basing on Carolyn Miller's (1994) concept of genre as a social action, concentrating mostly on the blog's communicative purpose and functions. Following the principles of the grounded theory (cf. Lonkila 1999) the team interpreted those blogs and created an empirical-conceptual typology which entailed following genres: diaries (subjective, self-referential discourse), reflection (subjective discourse on universal matters), criticism (subjective and expert discourse on general issues), information (objective facts), filter (gateway to the existing web content), advice (subjective and expert instructions on particular issues), modelling (serving as a role model for readers) and fictionality (description of fictional events). Weblogs in the sample were coded by three separate coders with 69% average pairwise percent agreement and Cohen's kappa of .622[1]. Such a moderate agreement could be attributed to the fact that the resulting genres are ideal types, and most of the actual blogs share features of more than one genre.

This subsequent study aims at supplementing this close-reading typology with a distant-reading perspective (Moretti 2013), based on selected tools for language processing and text clustering. We explore the style of those genres, adapting the definition proposed by Herrmann et al.: "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (2015:41). We chose this approach due to its stress on mixed methods, as we are combining linguistic and literary criteria of selecting style markers to discriminate between blog genres (Leech & Short 2007,57-58). Current research in the field of computational literary genre stylistics focuses on Most Frequent Words (e.g. Schöch and Pielström 2014; Jannidis & Lauer 2014) or functional linguistic categories (or both) (e.g. Allison et. al 2011). Yet, this study applies similar methods to emergent and uncategorised forms of writing. The quantitative methods are incorporated into the qualitative research workflow in order to create a productive feedback loop.

Corpus

The corpus of blogs was collected with the use of *BlogReader* - an extension of a corpus gathering system developed in CLARIN-PL (Oleksy et al. 2014) on a basis of open components: *jusText* and*Onion* (Pomikálek, 2011). From the initial set analysed by Maryl et al., 250 blogs were selected for processing as being long enough and included clean text (comments were omitted). We intentionally left out blogs with exceptionally large or small amount of text in order to balance the sample. The selected subcorpus includes: Diaries (44 blogs), Reflection (12), Criticism (73), Information (10), Filter (11), Advice (59), Modelling (24), Fictionality (5), and 10 'Unblogs', i.e. websites or portals using the label of blogs. Posts from the one blog were merged together into a single text document per a blog that was saved in the CCL corpus format (Broda et al., 2012).

Processing

We followed the blueprint of stylometry to find groups of blogs, e.g. (Burrows 2002), (Stamatatos, 2009) or (Eder, 2011). Blogs were described by feature vectors whose initial values were frequencies of the selected elements. They were next filtered or transformed. The transformed vectors were clustered into a number of groups that could be presented as automatically identified blog types or compared with the original types.

According to the criteria considered for the typology of blogs, we assumed that the interesting distinctions are not of semantic character. Thus we tried to define descriptive features that are not sensitive to the semantics of the blog contents. As a consequence, we have analysed features based on frequencies of lemmas, grammatical classes and sequences of grammatical classes. The brief description below will be elaborated in the presentation:

1. We have selected the 500 most frequent lemmas from the *Polish National Corpus* (Przepiórkowski et al., 2012) and in the series of experiments on the corpus of novels we reduced it to 212 lemmas that did not trigger semantic grouping (e.g. filtering out most of nouns and verbs).

2. Grammatical classes (as defined in the *Polish National Corpus* tagset) were recognised by *WCRFT* morpho-syntactic tagger (Radziszewski, 2013).

3. Features were defined and extracted with the help of the *Fextor* system (Broda et al., 2013).

4. Raw feature values were transformed by measures returning positive results for those features which contribute the significant amount of information to the document description. *SuperMatrix* system (Broda & Piasecki, 2013) for Distributional Semantics was applied during the transformation.

5. Similarity of the transformed vectors were computed by the cosine and ratio measures. The first is not sensitive to the differences in the document lengths that was the case of the analysed collection. The ratio as a heuristic measure that is aimed at comparing how much information is shared by the two vectors:

$$\mathrm{ratio}(V,U) = 2{*}\mathrm{sum}((Vi + Ui)/\max(Vi, Ui) - 1) / (\mathrm{length}(V) + \mathrm{length}(U))$$

6. Clustering was performed by the *Cluto* package for text data clustering (Zhao & Karypis 2005). In addition, *Stylo* package (Eder et al., 2013) for stylometry was used in experiments with visualisation of the possible blog clusters.

In order to understand the clusters better, most significant features for each cluster were identified and ranked. From several tests the Mann-Whitney U nonparametric test was chosen. For each feature its values in the documents of the given cluster were compared with its values in documents from the rest of the collection.

Experiments

We have performed several experiments that can be divided into three main groups:

1. *lexical level analysis*, based solely on the selected most frequent lemmas and punctuation marks and aimed at testing whether those properties can serve as a basis for automated identification of the blog types;

2. *lexico-syntactic level analysis* featured grammatical classes in combination with the lexical features of the lexical analysis in order to assess whether blog styles result in syntactic properties. On both levels we set the expected number of clusters to 20, in order to give algorithm more 'freedom';

3. extraction of significant features for the blog types with the help of the Mann-Whitney U nonparametric test.

Discussion

The generated groups represented relatively high average of clusters purity: 54%-60,4%, i.e. more than 50% blogs in a cluster are of same type. Entropy was higher than expected: 0.438-0.481, i.e. besides dominating types in clusters blogs of other types were scattered (especially smaller types). However, the obtained clusters did not match very well the qualitatively defined types. Lexical analysis combined with the ratio measure produced results that were closest to the qualitative types: entropy of 0.467 and 58% purity, see Figure 1. Yet, lexico-syntactic analysis (lexical features together with grammatical classes and bigrams) yielded better results: 0.438 of entropy and 60.4% of purity, see Figure 2. A slightly worse result: 0.481 of entropy and 54% of purity, was obtained with trigrams instead of bigrams - groups became too small and too specific.

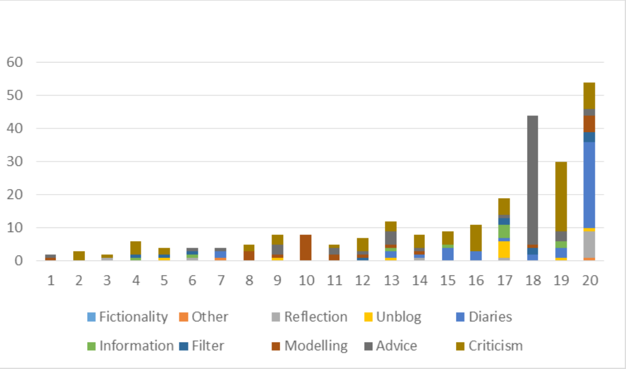


Figure 1. Results of the lexical analysis (features: 212 selected frequent lemmas, punctuation marks), PMI weighting, the ration similarity and, graph clustering algorithm from Cluto

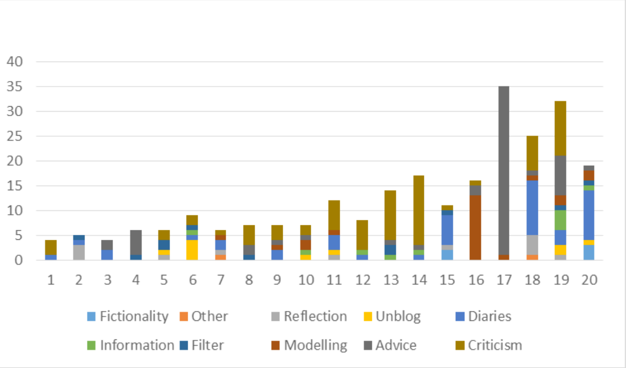


Figure 2. Results of the lexico-syntactic analysis (lexical features plus grammatical classes and bigrams), PMI weighting, ration similarity, graph clustering algorithm

Such genres as advice, criticism and, to certain extent, diaries and modelling were clustered together with others present in multiple clusters. It was caused by distinctive language features of those genres, especially of the advice, which employs instructional vocabulary, or criticism, due to its essayistic style with compound sentences and conjunctions reflecting logical reasoning. Diaries tend to use narrative language, whereas modelling blogs are clearly concentrated on expressing the author's self.

Those differences were further explored through the extraction of blog types' significant features with the use of Mann-Whitney U statistic. The results were in line with the definitions of classes, but provided more detailed information about the linguistic cues in those genres, some of which are presented in Table 1.

Conclusions

This study showed how close readings (literary interpretative practices) and distant readings (computational approaches to genre analysis) could be integrated in a non-topical analysis of the emerging genres. The novelty of the presented approach lies in the fact that we do not aim at assessing existing genres but rather at developing tools and procedures for the analysis and classification of new genres. The automated methods are used not only to verify the qualitative findings, but rather to enhance them by pointing towards the attributes which might have been overlooked by human coders who were able to read only a sample of each of 332 blogs. The aim is not to cluster texts automatically but rather to support human interpretation in an integrated research design.

Recurring problems with clustering genres other than advice could be attributed to the fact that individual blogs within one class may consists of posts which follow different genre conventions. Hence, further studies should explore the genre problem by comparing individual posts (rather than entire blogs) by different authors in order to find stylistic similarities.

Genre	Linguistic features
Advice	infinitive, passive adjectival participle, numerals, measurements (“about”, “large”, “small”)
Criticism	subjective vocabulary: „I”, „mine”; conjunctions pointing to logical reasoning, e.g. “if”, “that”, “given”, “hence”, “but”
Diaries	1st & 2nd person; vocabulary: “self”, “to be”; specific words and verb forms pointing out to a narrative: “certain”, “there”
Fictionality	past tense, 3rd person
Filter	punctuation, substantives
Information	impersonal verb forms, 3rd person
Modelling	interjections (e.g. “eh”), exclamation marks, 1st & 2nd person, vocabulary: “mine”, “thing”, “new”, “why”, “because”
Reflection	1st & 2nd person, vocabulary: “self”, “always”, “everything”

Table 1. Selected linguistic features of weblog genres (Mann-Whitney U)

Bibliography

Allison, S., Heuser, R., Jockers, M. L., Moretti, F. and Witmore, M. (2011). *Quantitative Formalism: An Experiment*, Pamphlet 1. Stanford Literary Lab.

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A. and Wardyński, A. (2012). *KPWr: Towards a Free Corpus of Polish*, *Proceedings of LREC'12*. Istanbul, Turkey.

Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R. and Wardyński, A. (2013). Fextor: A Feature Extraction Framework for Natural Language Processing: A Case Study in Word Sense Disambiguation, Relation Recognition and Anaphora Resolution. In Przepiórkowski, A., Piasecki, M., Jassem, K. and Fuglewicz, P. (eds), *Computational Linguistics. Applications*, volume 458 of Studies in Computational Intelligence, Berlin: Springer Verlag, pp. 41–62.

Broda, B. and Piasecki, M. (2013). Parallel, Massive Processing in SuperMatrix – a General Tool for Distributional Semantic Analysis of Corpora. *International Journal of Data Mining, Modelling and Management*, **5**(1): 1–19.

Burrows, J. F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of authorial fingerprint. *Studies in Polish Linguistics*, **6**: 99-114.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts*. University of Nebraska-Lincoln, NE, pp. 487-89.

Freelon, D. (2010). ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, **5**(1): 20-33.

Herring, S., Shedit, L. A., Writh, E. and Bonus, S. (2005). Weblogs as a bridging genre, *Information, Technology & People*, **18**(2): 142-71.

Herrmann, J. B., van Dalen-Oskam, K. and Schöch, Ch. (2015). Revisiting Style, a Key Concept in Literary Studies, *Journal of Literary Theory*, **1**(9): 25-52.

Jannidis, F. and Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In Erlin,M. and Tatlock, L. (eds), *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, New York: Camden House, pp. 29-54.

Leech, G. N. and Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose*. Harlow: Pearson Longman.

Lonkila, M. (1999). Grounded theory as an emerging paradigm for computer-assisted qualitative data analysis. In Kelle, U. (ed), *Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice*. London: Sage, pp. 41-51.

Maryl, M., Niewiadomski, K. and Kidawa, M. (2016 - forthcoming). Empirically Generated Typology of Weblog Genres. *CLCWeb: Comparative Literature and Culture*, **18**(2).

Miller, C. R. (1994). Genre as Social Action. In Freedman, A. and Medway, P. (eds), *Genre and the New Rhetoric*. London: Taylor & Francis, pp. 57-66.

Moretti, F. (2013). *Distant reading*. London: Verso.

Morrison, A. (2008). Blogs and Blogging: Text and Practice. In Siemens, R. and Schreibman, S. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell, pp. 369-87.

Oleksy, M., Kocoń, J., Maryl, M. and Piasecki, M. (2014). Linguistic analysis of weblog genres, *Practical Applications of Linguistic Corpora Conference*, PALC'14, Łódź.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD Thesis, Faculty of Informatics, Masaryk University, Brno. http://is.muni.cz/th/45523/fi_d/phdthesis.pdf (accessed 29 February 2016).

Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds). (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.

Radziszewski, A. (2013). A tiered CRF tagger for Polish. In Intelligent Tools for Building a Scientific Information Platform, *Studies in Computational Intelligence*, vol. 467, pp. 215–30.

Rybicki, J. and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, **26**(3): 315-21.

Schöch, Ch. and Pielström, S. (2014). Für eine computergestützte literarische Gattungsstilistik, *1. Jahrestagung der Digital Humanities im deutschprachigen Raum (DHd).*

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.

Zhao, Y. and Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis. Technical Report TR #01–40, Department of Computer Science, University of Minnesota, Minneapolis.

Zhao, Y. and Karypis, G. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, **10**(2): 141-68.

Notes

[1] The intercoder reliability was calculated with ReCal3, see (Freelon, 2010)

Wikidition: Towards A Multi-layer Network Model of Intertextuality

Alexander Mehler
mehler@em.uni-frankfurt.de
Goethe University, Germany

Benno Wagner
wagner@lit-wiss.uni-siegen.de
Beijing Institute of Technology, China

Rüdiger Gleim
Gleim@em.uni-frankfurt.de
Goethe University, Germany

Introduction

Current computational models of intertextuality run the risk of ignoring several desiderata: on the one hand, they mostly rely on single methods of quantifying text similarities. This includes syntagmatic models that look for shared vocabularies (unigram models) or (higher order) (*k*-skip-) *n*-grams (Guthrie et al., 2006). Such approaches disregard the two-level process of sign constitution according to which language-related, paradigmatic relations have to be distinguished from their text-related, syntagmatic counterparts (Hjelmslev, 1969; Miller et al., 1991; Raible, 1981) where the former require language models of the sort of neural networks (Mikolov et al., 2013), topic models (Blei et al., 2007) or related approaches in the area of latent semantic analysis (cf., e.g., (Paaß et al., 2004)). On the other hand, computational models should enable scholars to revise their computations. The reason is the remarkably high error rate produced by statistical models even in cases that are supposed to be as “simple” as automatic *pre*-processing. Thus, scholars need efficient means to make numerous corrections and additions to automatic computations. Otherwise, the computations will be hardly acceptable as scientific data in the humanities (Thaller, 2014).

	$\{\textbf{hypotext}_i \mid i = 1\}$	$\{\textbf{hypotext}_i \mid i > 1\}$	$\{\textbf{hypotext}_i \mid i \gg 1\}$
$\{\textbf{hypertext}_i \mid i = 1\}$	(1)	(2)	(3)
$\{\textbf{hypertext}_i \mid i > 1\}$	(4)	(5)	(6)
$\{\textbf{hypertext}_i \mid i \gg 1\}$	(7)	(8)	(9)

Table 1: Nine scenarios of generating Wikiditions out of corpora of (referring) literary (hyper-)texts and their (referred) hypotexts. Examples: (1) Kafka's "Bericht für eine Akademie" (in the role of a hypertext) versus Hauff's "Der Affe als Mensch" (in the role of a hypotext); (2) Kafka's "Bericht für eine Akademie" versus all "Affentexte" (Borgards, 2012) since the end of the 18th Century (including works of, e.g., Hauff, E. T. A. Hoffmann, Flaubert etc.); (3) Kafka's "Beim Bau der Chinesischen Mauer" versus the "Prager Tagblatt" from August 1914 to March 1917; (4) Kafka's "Oeuvre" versus Nietzsche's "Geburt der Tragödie aus dem Geiste der Musik"; (5) a selection of Kafka's "Oeuvre" versus a selection of Nietzsche's "Oeuvre"; (6) Kafka's "Oeuvre" versus a newspaper corpus (e.g., sampled from the "Prager Tagblatt"); (7) the complete works of several authors versus a single hypotext (e.g., Goethe's "Faust"); (8) the complete works of several authors versus a corpus of "Faust" texts; (9) the complete works of several German authors versus the complete works of several French authors.

This paper presents *Wikidition* as a *Literary Memory Information System* (LiMeS) to address these desiderata. It allows for the automatic generation of online editions of text corpora. This includes literary texts in the role of (referring) *hypertexts* (Genette, 1993) in relation to candidate (referred) hypotexts by exploring their *intra- and intertextual relations* – see Table 1 for nine related research scenarios. In order to explore, annotate and display such relations, Wikidition computes multi-layer networks that account for the multi-resolution of linguistic relations – on the side of the hypo- and the hypertexts. The reason is that hypertextual relations (in the sense of Genette) (that occur in the form of transformations, imitations or any mixture thereof) may be manifested on the lexical, sentential or the textual level (including whole paragraphs or even larger subtexts). As a consequence, Wikidition spans lexical, sentential and textual networks that allow for browsing along the constituency relations of words in relation to sentences, sentences in relation to texts etc. In this multi-layer network model, intrarelational links (of words, sentences or texts) are represented together with interrelational links that combine units of different layers. Figure 1 shows the range of sign relations that are mapped. To this end, Wikidition combines a multitude of text similarity measures (beyond *n*-grams) for automatically linking lexical, sentential and textual units regarding their (1) syntagmatic (e.g., syntactic) and (2) paradigmatic use. We call this two-level task *linkification*.

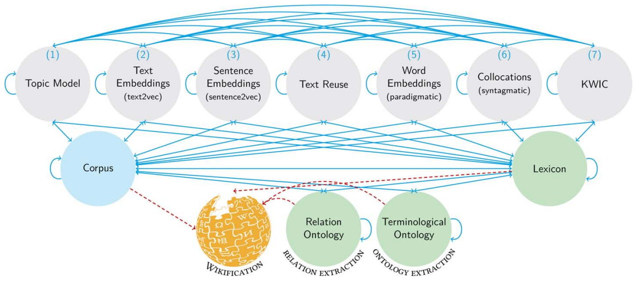


Figure 1: Sign relations that are automatically explored and annotated by Wikidition (Mehler et al., 2016): on the level of words (Module (5) – paradigmatic –, (6) and (7) – both syntagmatic), on the level of sentences (Module (3) – paradigmatic – and (4) – syntagmatic) and on the level of texts (Module (1) and (2) – both paradigmatic). Wikidition additionally includes a component for wikification (i.e., for linking occurrences of concepts to articles in Wikipedia (Mihalcea et al., 2007)) and especially for automatically inducing lexica out of input corpora (i.e., for linkification). Arcs denote links explored by Wikidition; reflexive arcs denote intrarelational (i.e., purely lexical, sentential or textual) links. In this way, intra- and interrelational links are maintained by the same information system.

Beyond linkification, Wikidition contains a module for automatic *lexiconisation* (see Figure 2). It extracts lexica from input corpora to map author specific vocabularies as subsets of the corresponding reference language. Input corpora (currently in English, German or Latin) are given as plain text that first are automatically preprocessed; the resulting wikiditions are mapped onto separate URLs to be accessible as self-contained wikis. By means of lexiconisation, research questions of the following sort can be addressed: *What kind of German does Franz Kafka write? (E.g., Prager Deutsch.) What terminologies does Franz Kafka use in "In der Strafkolonie"? (E.g., engineering terminology.) How does his German depart from the underlying reference language?* Since texts are not necessarily monolingual (because of using citations, translations, loan words, verbal expressions etc.), the same procedure can be applied by simultaneously looking at all foreign languages being manifested in the texts under consideration (right side of Figure 2).

To this end, Wikidition distinguishes three levels of lexical resolution: superlemmas (e.g. German *Tätigkeit*), lemmas (e.g., *Thätigkeit*) and syntactic words (e.g., *Thätigkeiten* (*nominative*, *plural*)) as featured sign-like manifestations of lemmas (lower part of the figure). Note that this model diverges from the majority of computational models to textual data which start from tokens as manifestations of wordforms (referred to as types) and which, therefore, disregard the meaning-side of lexical units. Based on linkification and lexiconisation, Wikidition does not only allow for traversing input corpora on different (lexical, sentential and textual) levels. Rather, the readers of a Wikidition can also study the vocabulary of single authors on several levels of resolution: starting from the level of superlemmas via the level of lemmas down to the level of syntactic words and wordforms (see Figure 2).

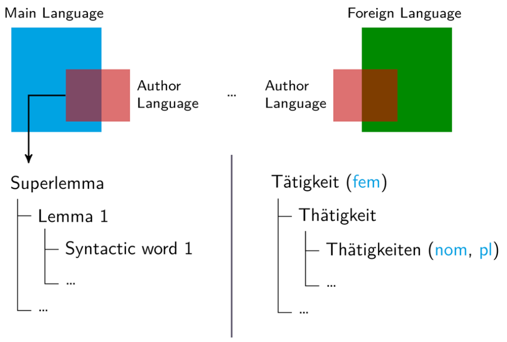


Figure 2: Left side: schematic depiction (red) of the vocabulary of an author (e.g., Franz Kafka) as manifested within Wikidition's input text(s) (e.g., "In der Strafkolonie") as mainly overlapping with the vocabulary of the corresponding reference language (e.g., German).

	Human Close Reading	Distant Reading (Moretti 2013)	Machine Reading (Etzioni 2007)	Machine Close Reading
Research object	$\{T_1 \dots, T_m \mid X_n\}$ $m \rightarrow 1, n > 1$	$\{T_1, \dots, T_m\}$ $m \rightarrow \infty$	$\{T_1, \dots, T_m\}$ $m \rightarrow \infty$	$\{T_1 \dots, T_m \mid X_n\}$ $m \ll n, n \rightarrow \infty$
Quantity of data	small data	big data	small → big	big → small
Quantification	implicit	machine-based	machine-based	twofold
Interpretation	human-based	human-based	machine-based	human based
Research focus	understanding	hidden laws	understanding	hypotheses testing
Resources	human mind	corpus + HM	corpus + SW	corpus+HM+SW

Table 2: Notions of human, computer-supported and machine-based reading. Wikidition addressesmachine close reading by integrating semantic web (SW) resources and the human mind (HM) (asthe ultima ratio of interpreting its computations). T_1, ..., T_m span the input corpus of m (hyper-)texts; X_n denotes the contextualizing corpus of hypotexts of size n that is explicitly consulted by the readingprocess. Machine close reading is similar to human reading in that it focuses on small, rather than bigdata.

While the linkification component of Wikidition relates to principles of WikiSource and Wikipedia, the Wiktionary project is addressed by its lexiconisation module. Wikidition uses numerous computational methods for providing interoperability and extensibility of the resulting editions according to the wiki principle. In this way, the dissemination of computer-based methods is supported even across the borders of digital humanities in that scholars are enabled to make their own exploratory analyses. However, Wikidition does not address a big data scenario in support of distant reading (Moretti, 2013), nor does it aim at emulating human reading in the sense of machine reading (Etzioni, 2007). Rather, Wikidition addresses what we call *machine close reading* in that it aims at massively supporting the process of (scientific or literary) reading by means of computational methods (see Table 2).

Evaluation

We exemplify Wikidition by example of three pairs of text. Regarding the layers of lemmas and sentences, Table 3 shows that Wikidition generates extremely sparse networks (whose cohesion is below 1%) of high cluster values and short average geodesic distances in conjunction with largest connected components that comprise almost all lexical and sentential nodes. In this example, we compute paradigmatic associations among words by means of word2vec (Mikolov et al., 2013) while sentence similarities are computed by means of the centroids of the embeddings of their lexical constituents. Networks are filtered by focusing on the first three most similar neighbors of each node – obviously, this does not interfere with the small-world topology of the networks. Each pair of texts is additionally described regarding the subnetwork of syntactic words and sentences. This is done to account for the impact of inflection on networking. As a result, the networks are thinned out (cohesion is now at most 0.5%), but neither the sizes of the largest connected components nor the cluster and distances values are affected considerably. Obviously, differentiation leads to sparseness, but in a sense that the general topology is retained. By focusing on a single level of resolution (e.g., paradigmatic relations among words), sub-networks are generated that fit into what is known about universal laws of complex linguistic networks (Mehler, 2008). See (Mehler et al., 2016) for additional evaluations of Wikidition.

Edition	**#nodes**	**#links**	**C**	**lcc**	**l**	**coh**
Kafka // Nietzsche	1624	10391	0,27	1	3,31	0,008
	2401	13644	0,34	1	3,56	0,005
Kafka // Rathenau	1749	10782	0,28	1	3,36	0,007
	2473	13609	0,36	1	3,60	0,004
Kafka // Rauchberg	4034	30951	0,25	0,999	3,35	0,004
	6830	44369	0,31	0,999	3,62	0,002

Table 3: Wikiditions of three text pairs (Kafka: Beim Bau der Chinesischen Mauer // Nietzsche: Die Zeit der Zyklopen-Bauten; Kafka: Ein Bericht für eine Akademie // Rathenau: Höre, Israel; Kafka: In der Strafkolonie // Rauchberg: Statistische Technik) compared by their cluster value c, the average geodesic distance of their nodes l, the fraction of nodes in their largest connected components lcc and by their cohesion coh (the number of links in relation to all possible links). First line of each text pair: nodes comprise lemmas and sentences; second line of each pair: nodes comprise syntactic words and sentences. Networking is conditioned by the operative preprocessor (Eger et al., 2016).

Conclusion

We presented Wikidition as a framework for exploring intra- and intertextual relations. Wikidition combines machine learning with principles of several wiki-based projects (Wikipedia, WikiSource and Wiktionary) to generate multi-layer networks from input corpora by integrating syntagmatic and paradigmatic relations on the lexical, sentential and the textual level. Our approach addresses intra- and inter-level networking in a single framework while adhering to laws of networking as being explored by complex network theory. In this way, input corpora get traversable in line with both empirical findings about characteristics of linguistic networks and the multi-resolution of sign relations whose space complexity is preferably reduced. Currently, Wikidition exists as a prototype that is further-developed by means of several edition projects in order to be finally made available as open source software. Wikidition is open for the cooperative development of digital editions.

Bibliography

Eger, S. and Gleim, R. and Mehler, A. (2016). Lemmatization and Morphological Tagging in German and Latin: A comparison and a survey of the state-of-the-art. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Etzioni, O. (2007). Machine reading of web text. In *Proceedings of the 4th International Conference on Knowledge Capture, K-CAP '07*, pp. 1–4.

Genette, G. (1993). Palimpseste: Die Literatur auf zweiter Stufe. Suhrkamp, Frankfurt am Main.

Guthrie, D., Allison, B., Liu, W., Guthrie, L. and Wilks, Y. (2006). A closer look at skip-gram modelling.

Hjelmslev, L. (1969). Prolegomena to a Theory of Language. University of Wisconsin Press, Madison.

Kafka, F. (1916). Die Verwandlung. Kurt Wolff Verlag, Leipzig.

Kafka, F. (1919). In der Strafkolonie. Kurt Wolff Verlag, Leipzig.

Mehler, A. (2008). Large text networks as an object of corpus linguistic studies. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook of the Science of Language and Society*. De Gruyter, Berlin/New York, pp. 328–82.

Mehler, A., Gleim, R., vor der Brück, T., Hemati, W., Uslu, T. and Eger, S. (2016). "Wikidition: Automatic Lexiconization and Linkification of Text Corpora," *Information Technology*.

Mihalcea, R. and Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07*. New York, NY, USA. ACM, pp. 233–42.

Mikolov, T., Yih, W. and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, (eds), *Proceedings of NAACL 2013*. The Association for Computational Linguistics, pp. 746–51.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, **6**(1): 1–28.

Moretti, F. (2013). Distant Reading. *Verso*.

Paaß, G., Kindermann, J. and Leopold, E. (2004). Learning prototype ontologies by hierarchical latent semantic analysis. In Andreas Abecker, Steffen Bickel, Ulf Brefeld, Isabel Drost, Nicola Henze, Olaf Herden, Mirjam Minor, Tobias Scheffer, Ljiljana Stojanovic, and Stephan Weibelzahl, (eds), *LWA 2004: Lernen – Wissensentdeckung – Adaptivität*. Humbold-Universität Berlin, pp. 193–205.

Raible, W. (1981). Von der Allgegenwart des Gegensinns (und einiger anderer Relationen). Strategien zur Einordnung semantischer Informationen. *Zeitschrift für romanische Philologie*, **97**(1-2): 1–40.

Rauchberg, H. (1890). Statistische Technik. Deutsche Statistische Gesellschaft, 1.

Thaller, M. (2014). The humanities are about research, first and foremost; their interaction with computer science should be too. *Dagstuhl Reports*, **4**(7): 108–10.

Mapping the Bentham Corpus

Estelle Tieberghien
estelle.tieberghien@gmail.com
LATTICE-CNRS, France

Pablo Ruiz Fabo
pablo.ruiz.fabo@ens.fr
LATTICE-CNRS, France

Frédérique Mélanie-Bécquet
frederique.melanie@ens.fr
LATTICE-CNRS, France

Thierry Poibeau
thierry.poibeau@ens.fr
LATTICE-CNRS, France

Tim Causer
t.causer@ucl.ac.uk
University College London, UK

Melissa Terras
m.terras@ucl.ac.uk
University College London, UK

Introduction

University College London (UCL) owns a large corpus of the philosopher and social reformer Jeremy Bentham (1748-1832). Until recently, these papers were for the most part untranscribed, so that very few people had access to the corpus to evaluate its content and its value. The corpus is now being digitized and transcribed thanks to a large number of volunteers recruited through a crowd-sourcing initiative called Transcribe Bentham (Causer and Terras, 2014a, 2014b).

The problem researchers are facing with such a corpus is clear: how to access the content, how to structure these 30,000 files, and how to get relevant access to this mass of data? Our goal has thus been to produce an automatic analysis procedure aiming at providing a general characterization of the content of the corpus. We are more specifically interested in identifying the main topics and their structure so as to provide meaningful static and dynamic representations of their evolution over time.

Comparison with other works

The exploration of large corpora in the Humanities is a known problem for today's scholars. For example, the recent PoliInformatics challenge addressed the issue by promoting a framework to develop new and original research in text-rich domains (the project focused on political science but can be extended to any sub-field within the Humanities).

Specific experiments have recently been done in the field of philosophy, but they mainly concern the analysis of metadata, like indexes or references (Lamarra and Tardella, 2014; Sula and Dean, 2014). Different experiments have nevertheless involved an exploration of large amounts of textual data (see e.g. Diesner and Carley, 2005 on the Enron corpus) with relevant visualization interfaces (Yanhua et al., 2009).

In this paper, we propose to explore more advanced natural language processing techniques to extract keywords and filter them according to an external ontology, so as to obtain a more relevant indexing of the documents before visualization. We also explore dynamic representations, which were not addressed in the above-mentioned studies.

Corpus exploration strategy

The Text analysis module

Different scripts have been developed to filter the corpus[1]. Then documents are assigned a date whenever possible: Since the corpus mostly contains notes and letters, the first date mentioned in the document often refers to the date of the document's composition (even if this assumption is of course not always true). A large number of documents cannot be assigned a date and are thus not used for the dynamic analysis of the corpus.

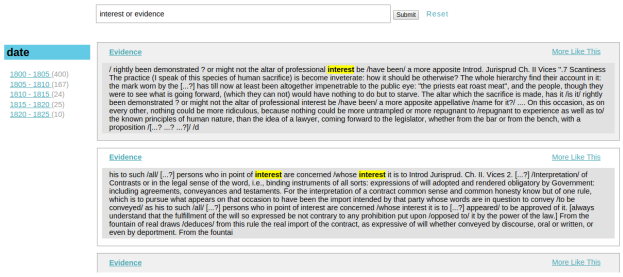
To index the corpus and identify meaningful concepts, we first tried to directly extract relevant keywords from the texts. However, traditional techniques like the use of tf-idf (Salton et al., 1983) and c-value (Frantzi et al., 2000) do not seem very efficient in our case. This is not too surprising: it is well known that texts are too ambiguous to provide a sound basis for a direct semantic extraction. Surface variations, the use of synonyms and hyponyms, linguistic ambiguity and other factors constitute strong obstacles for the task. We thus decided to use natural language processing techniques that provide relevant tools to overcome some of these limitations. The tools we employed are either web-based or possible to execute on a personal computer with average specs.

We tried to refine concept extraction by confronting the text with an external, structured database. We used DBpedia (Auer et al., 2007) as a source of structured knowledge (DBpedia is a database made of information extracted from Wikipedia). DBpedia is not a specialized source of information but this guarantees that the approach is not domain or author specific and could be easily used for other corpora. We then used the DBpedia Spotlight Web Service (Mendes et al., 2012; Daiber et al., 2013) to make the connexion between the corpus and DBpedia concepts. This leads to a much more fine grained and relevant analysis than possible with an entirely data-driven keyword extraction.

Based on the outputs of Spotlight, only concepts that occurred at least 100 times, and with a confidence value of at least 0.1 were kept. Spotlight outputs a confidence value between 0 and 1 for each annotation; a 0.1 threshold removes clearly unreliable annotations while maintaining good coverage. Tagging the full corpus with Spotlight (ca. 30,000 documents) took over 24 hours. We called the Spotlight service one document at a time; parallelizing the process can decrease processing time.

The visualization module

Once relevant concepts are identified, one wants to produce relevant text representations so as to provide a usable interface to end users. We present here three different kinds of interfaces that show the possible exploitation of the analysis described above.



interest or evidence

1800 - 1805 (400)
1805 - 1810 (167)
1810 - 1815 (24)
1815 - 1820 (25)
1820 - 1825 (10)

Evidence

/ rightly been demonstrated ? or might not the altar of professional interest be /have been/ a more apposite Introd. Jurisprud Ch. II Vices ".7 Scantiness The practice (I speak of this species of human sacrifice) is become inveterate: how it should be otherwise? The whole hierarchy find their account in it: the mark worn by the [...?] has till now at least been altogether impenetrable to the public eye: "the priests eat roast meat", and the people, though they were to see what is going forward, (which they can not) would have nothing to do but to starve. The altar which the sacrifice is made, has it /is it/ rightly been demonstrated ? or might not the altar of professional interest be /have been/ a more apposite appellative /name for it?/ .... On this occasion, as on every other, nothing could be more ridiculous, because nothing could be more untrampled or more repugnant to /repugnant to experience as well as to/ the known principles of human nature, than the idea of a lawyer, coming forward to the legislator, whether from the bar or from the bench, with a proposition /[...? ...? ...?]/ /d

Evidence

his to such /all/ [...?] persons who in point of interest are concerned /whose interest it is to Introd Jurisprud. Ch. II. Vices 2. [...?] /Interpretation/ of Contrasts or in the legal sense of the word, i.e., binding instruments of all sorts: expressions of will adopted and rendered obligatory by Government: including agreements, conveyances and testaments. For the interpretation of a contract common sense and common honesty know but of one rule, which is to pursue what appears on that occasion to have been the import intended by that party whose words are in question to convey /to be conveyed/ as his to such /all/ [...?] persons who in point of interest are concerned /whose interest it is to [...?] appeared/ to be approved of it. [always understand that the fulfillment of the will so expressed be not contrary to any prohibition put upon /opposed to/ it by the power of the law.] From the fountain of real draws /deduces/ from this rule the real import of the contract, as expressive of will whether conveyed by discourse, oral or written, or even by deportment. From the fountai

Evidence

Figure 1: Search interface: users can search via year extracted from the text, which in most cases is the year of writing, allowing users to see texts (especially correspondence) in chronological order.

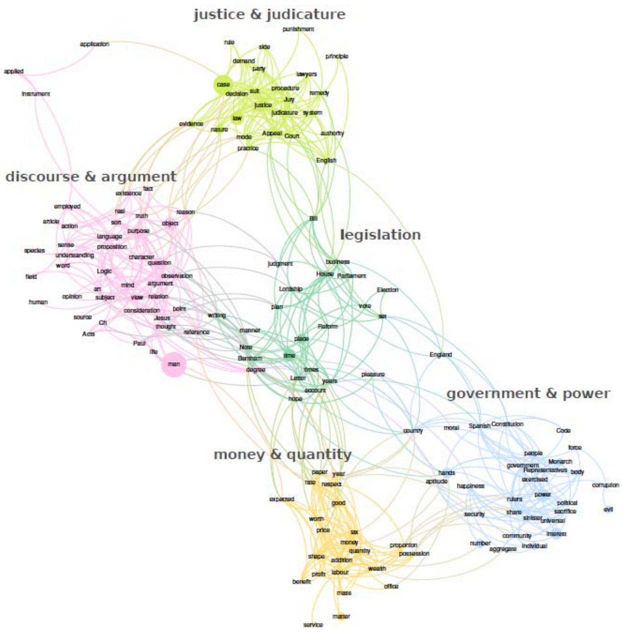


Figure 2: the main topics addressed in the corpus, based on clusters of concepts, showing the main concerns of Bentham's writings, which map closely onto established research areas in Bentham studies. The network was produced by Cortext; colours and fonts were reformatted in Gephi based on Cortext's gexf-format export[4]

The corpus is first indexed in a Solr search index[2] and accessible through a graphical end-user interface. It is possible to query the corpus by date, using Solr's faceted search functions[3] (see figure 1).

It is also possible to cluster together related keywords, so as to get access to homogeneous sub-parts of the corpora representing specific subfields of Bentham's activity (see figure 2).

Dynamic maps are also possible, to see for example the evolution of the different topics addressed in the corpus over time (see figure 3).

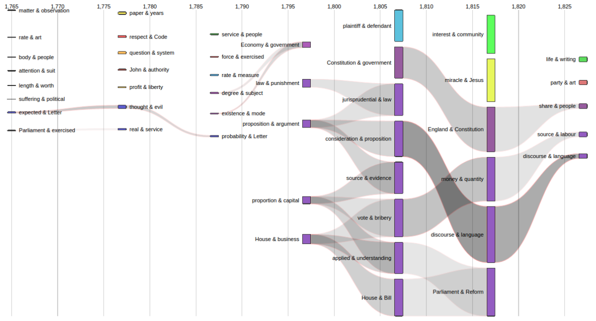


Figure 3: A dynamic view of the corpus, computed with the Cortext plarform *(tubes layout)*, with the evolution of the main topics addressed over time

Scholarly benefits of these tools for the Transcribe Bentham project

Since 1958, UCL's Bentham Project has been producing the new, critical edition of the "Collected Works of Jeremy Bentham". The edition is expected to run to some eighty volumes, the thirty-third of which has recently been sent to the press. The "Collected Works" is based upon texts, which Bentham published during his lifetime, and unpublished texts, which exist in manuscript. It is a major task: UCL's Bentham Papers runs to some 75,000 manuscript pages, while the British Library's has a further 25,000 or so pages. About 40,000 pages have been transcribed to date and, while UCL's award-winning 'Transcribe Bentham' initiative has helped to significantly increase the pace of transcription, a great deal more work needs to be done.

The first task in producing a volume of the "Collected Works" based upon texts in manuscript is to identify all the relevant pages. Bentham Project editorial staff use the Bentham Papers Database Catalogue, which indexes the manuscript collection by sixteen headings, including date, main heading, subject heading(s), author(s), and so forth. It is, however, entirely possible to miss relevant manuscripts using this method. The subject maps produced for this research promise to complement traditional Bentham Project methods; for instance, Bentham's work on political economy encompasses topics as varied as income tax to colonisation, and the subject maps will make it more straightforward to investigate the nexus between these, and other, subjects.

The dynamic corpus view, showing the evolution of topics addressed over time, could also prove useful in editorial work as can be shown in two examples. First, an editor at the Bentham Project is currently working on Bentham's writings on convict transportation, though there is some confusion over when exactly Bentham first broached the topic. The dynamic corpus view could help to clear up whether it was only around 1802 when Bentham wrote about transportation, or if he had investigated the subject in any great detail during the 1790s. Second, Bentham became more radical as he aged, and several Bentham scholars have sought to identify the point at which Bentham abandoned his earlier conservatism and 'converted' to political radicalism, and representative democracy; an analysis of Bentham's language at the turn of the century would be instructive in helping clarify this matter.

Conclusion

In this paper, we have presented a first attempt to give a relevant access to a large interdisciplinary corpus in the domain of philosophy, law and history. We have shown that using tools in concept clustering and visualization can provide an alternative way to navigate large-scale corpora, and confirm and visualise scholarly approaches to large scale textual corpora. Exploring how these tools can be effectively used with a corpus such as Bentham's indicates these methods are applicable to other sources as well.

In the near future, we are planning to refine the linguistic analysis in order to give better representations of the textual content of the corpus. We are also planning experiments with end-user to evaluate in more details the solution and the visualisation techniques used so far in this project.

Acknowledgements

We want to thank IDEX PSL (Paris Sciences et Lettres, ref. ANR-10-IDEX-0001-02 PSL) as well as the labex TransferS (laboratoire d'excellence, ANR-10-LABX-0099) for supporting this research. Lastly, we would like to thank the reviewers for their insightful comments on the paper.

Bibliography

Auer, S. (2007). DBpedia: A nucleus for a web of open data. *The Semantic Web*. Berlin Heidelberg: Springer.

Causer, T. and Terras, M. (2014a). Many hands make light work. Many hands together make merry work: *Transcribe Bentham* and crowdsourcing manuscript collections. In Ridge, M. *Crowdsourcing Our Cultural Heritage*, Farnham: Ashgate.

Causer, T. and Terras, M. (2014b). Crowdsourcing Bentham: Beyond the Traditional Boundaries of Academic History, *International Journal of Humanities and Arts Computing*, **8**(1): 46-64.

Daiber, J., Jakob, M., Hokamp, C. and Mendes, P. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, Graz.

Diesner, J. and Carley, K. (2005). Exploration of Communication Networks from the Enron Email Corpus. *Proceedings of SIAM International Conference on Data Mining, Newport Beach: Workshop on Link Analysis, Counterterrorism and Security*, pp. 3- 14,.

Frantzi, K., Ananiadou, S. and Mima H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, **3**(2): 115.

Lamarra, A. and Tardella, M. (2014). Theophilo. A prootype for a thesaurus of philosophy. Lausanne: *Digital Humanities 2014*.

Mendes, P., Daiber, J., Rajapakse, R., Sasaki, F. and Bizer, C. (2012). Evaluating the Impact of Phrase Recognition on Concept Tagging. Istanbul: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2012).*

Salton, G., Fox, E. and Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM 26.11*, pp. 1022-36.

Sula, C. and Dean, W. (2014). Visualization of Historical Knowledge structures: an Analysis of the Bibliography of Philosophy. Lausanne: *Digital Humanities 2014*.

Yanhua, C., Lijun, W., Ming, D. and Jing, H. (2009). Exemplar-based Visualization of Large Document Corpus. *IEEE Transactions on Visualization and Computer Graphics* (InfoVis2009), **15**(6): 1161-68.

Notes

[1] For example, Bentham sometimes used French in his correspondence and these texts are eliminated via automatic language detection, since we focus on English in this experiment.

[2] https://lucene.apache.org/solr/

[3] https://wiki.apache.org/solr/SolrFacetingOverview

[4] https://gephi.github.io/

Contextualizing Receptions of World Literature by Mining Multilingual Wikipedias

Ben Miller
miller@gsu.edu
Georgia State University, United States of America

Cindy Berger
cberger@student.gsu.edu
Georgia State University, United States of America

Sayan Bhattacharyya
sayan@illinois.edu
University of Illinois at Urbana Champaign, United States of America

Tommaso Caselli
tcaselli@vu.nl
VU University, Amsterdam, the Netherlands

David Kelman
dkelman@Exchange.FULLERTON.EDU
California State University, Fullerton, United States of America

Jennifer Olive
jolive1@gsu.edu
Georgia State University, United States of America

Jay Rajiva
jrajiva@gsu.edu
Georgia State University, United States of America

Reading in translation is an impoverishment, not so much because of the fluctuating quality of a translation or the loss of a perceived 'original work' but because of the elision of sociolinguistic context and the difficulty in conveying that lost context to readers. That world literature is taught almost always in translation at universities in America and elsewhere (a situation driven by the low capacity for foreign language instruction at the university level and the broad linguistic reach of world literature courses) compounds this problem. Of the estimated 20.2m American undergraduates in 2015, an estimated 1.6m are enrolled in foreign language courses, and, historically, only 17% of those reach the higher levels of proficiency necessary to read literature (Goldberg et al., 2013). The problem of American monolingualism has even affected the publishing industry; as noted translation theorist Laurence Venuti argued (1998), there are more translations of English texts into other languages than ever before while fewer foreign texts are translated into English. This trend means that American students are even less likely to feel comfortable with literature in translation than students from non-English speaking countries.

However, conversations about literary summarization, a core practice of critical reading, take place globally and are preserved in the background of pages on every national language Wikipedia. As each language community summarizes their translations of works of world literature in the form of Wikipedia articles, they generate for every crowdsourced entry a discussion page and a history of that page; together, they reveal that literary work's history of reception for a reading community in a given language. By striving to synthesize an authoritative, peer-reviewed summary of text native to or translated into that community's language, each group highlights their concerns, thought processes, and the challenges posed by a given work. The goal is to make these differences and conversations more visible to readers through techniques from natural language processing for automatic translation, topic modeling, and the visualization of topic models; in so doing, we aim to develop a method by which the digital humanities can address a core problem in comparative and world literature.

Machine translated Wikipedia discussions can help reveal to monolingual audiences the degree to which cultural pragmatics influence the reception of key (and popular) works of world literature. Although flawed, automatic translation has been found in some languages to be comparable with human translation, at least in regards to cohesion and formality (Li et al., 2014). Presenting parallel national-language conversations – such as the national language conversations about topics like translating the title for Camus' *L'Étranger / The Stranger / Der Fremde/ The Foreigner* and a visualization of the change-over-time of Goethe's *Faust I* summarizations in English, German, and Spanish – would help demonstrate the practical reception of a work in a language community. While work has been done on multi-lingual topic models (Ni et al., 2009), our research assumes that there will be both alignment and misalignment of topics across the various languages of a work; as such, our project resists the urge to normalize those topics into one category on the basis of an imperfect vector model of semantic similarity. Furthermore, experiments on iterative summarizations, such as elements of *The Tale of Genji*, demonstrate how even basic tasks in literary scholarship contain cultural dimensions and can thus reveal strong cultural patterns and biases held by different populations (Kashima, 2000). Along with the works mentioned above, this research explores the Wikipedia conversations around J.D. Salinger's *Catcher in the Rye / Der Fänger im Roggen / El guardián entre el centeno o El cazador oculto / L'Attrape-cœurs / Il giovane Holden* and Homer/Omero's *The Odyssey / Die Odyssee / Odisea / L'Odyssée / Odissea* across English, German, Spanish, French, and Italian Wikipedia pages.

In an increasingly digitized cultural landscape, the pedagogical use of Wikipedia and similar platforms has gained a great deal of traction in certain fields. While it is often derided as a dubious source for information, Wikipedia has proven to be a successful arena for instructing students on the distillation of information, writing in the public sphere, and collaborative writing (Purdy, 2009; Vetter, 2014; Sweeney, 2012). As a multilingual space, Wikipedia has offered many scholars the opportunity to examine the ways in which the cross-cultural sharing of information takes place (Nothman et al, 2013; Filatova, 2012). In some instances, Wikipedia specifically has been used as a place for the comparison of knowledge and representation across cultural and geographic divides (Callahan and Herring, 2011). These scholars provide a framework for examining the cross-linguistic aspects of Wikipedia in order to highlight cultural differences and deconstruct colonial power structures that privilege the English language (Ensslin, 2011). In the available scholarship, the use of Wikipedia in teaching literature has been largely ignored; only a few studies exist and those mainly address literary studies' reticence to incorporate Wikis into pedagogy (Bayliss, 2013). However, quite a few studies suggest that Wikipedia can occupy a distinct operational space in the university that supplements established pedagogical spaces and practices (Gorard and Selwyn, 2015; Knight and Pyke, 2012).

Our computational framework follows these possibilities to supplement two established methods of teaching translation in a world literature classroom. The first method consists of placing a work of literature in translation – a passage from a novel, for example – in relation to the same work in its original language. The hope is that students can analyze the differences between the original and the work in translation and, thereby, understand the way the literary text undergoes a 'new life' in translation. This method not only assumes a sophisticated knowledge of foreign languages among the majority of students but also brackets the question of how these works are read in the original language by native-language readers. The second method juxtaposes several translations of the same work into English. This has the advantage of not assuming knowledge of a foreign language. For example, students might read various translations of the 19th-century French poet Baudelaire in English, starting from English translations from the late 19th century and continuing through translations that have been published in the last twenty years. The advantage of this method is clear: students can easily see, in their own language, the way translation can change the meaning of a poem as they read it from one translation to another. While this method successfully bypasses the problem of foreign language competency, it amplifies the second problem: the student is even more divorced from the source-language context since all emphasis is placed on the way *English speakers* respond to a work of literature.

New methods, therefore, are needed to expose foreign texts and foreign contexts. Using computational methods to enter the themes and arguments about specific texts in other languages precludes some of the need for high-level competency in a foreign language in order to understand the debates about world literature in a non-American – and especially non-English – context. Automatic translation and topic modeling facilitate encounters with Wikipedia 'talk' pages in other languages. The rough output of these automated and annotated procedures preserves some of the estrangement of working across languages as the translation is rough, clearly communicating its nature as translation — and therefore serves its purpose without fully replacing the source text. With such experiments, our project broadly addresses two questions: how different language populations create summaries that are culturally distinct, and how these differences can be folded back into meaningful encounters for readers of works in translation. The overall method for this project is to identify the subset of well-documented and significant works, mine the relevant Wikipedia entries and conversations, and develop preliminary code to identify the linguistic features of the entry (e.g. topics, use of modals, noun density, phraseological structures, complexity measures, etc.) and the nodal points of the crowd's conversation that yielded the page. As an example, consider the discrepancies across the Italian and English discussion pages of *The Odyssey / Odissea* as represented by ten 5- to 10-term topic models.

Terms, Italian Discussion Page	Topic Name
odissea originali palla contributi registrami	Contributions and Registration
wikipedia forum migliorarla posto figlio arrivassero tenter tentare importanza	Need for More Expert Contributors
ulisse niente aggiunto dante manchi riferimento elenco procedo cielo parte pietose condizioni	References to Dante
poseidone divina generale voglia	Unclear
commedia incontra quesiti accecato ostacola competenti voce	Other Web Sources
ripristino aggiungerei siti porre partenza traducendo manna materia versione	Older Versions are Better than Newer Versions
dir notizie rete evidente polifemo	Unclear
inglese qualcuno pensa piccola inferno appositi pagina discutere film serve decise	Using English Wikipedia to Backfill Italian
esattamente pecca opere fin tema passo potere utenti	Reflections on Contributors
pare so troia attendo basandomi mesi trova provvisorio	Reflections on What to Write

odyssey homer work titles guideline searching works section promotional directed	Editing Guidelines and Work Title
february page common dab crazynas people epic redirect iliad dictionary	Genre and Other Works
talk odysseus december journey extant account term proposal giu september	The Journey
word article wp called style subject similar adventures locations Toronto	Locations in the Story and Writing Style
utc play topic primary part poem link note odisseus july	Poetry
title article refer word davidiad edited cynwolfe don titles voyage	Voyages
edit western preceding map request www apology talk zcc suggest	Mapping the Journey
medea akhilleus euripides noun don current literature click crazynas english	Characters
utc april added source staged meant argument written fall Cite	Sourcing and Staging
talk comment unsigned university musical oldest review tedickey tomb point	Ceremony

Table 1. Named Topics Models from English and Italian Discussion Pages of Omero/Homer’s *The Odyssey / Odissea*

What comes across in this comparison is the somewhat different concerns of the two reading communities. The Italian discussion reflects concerns with the expertise of the editing community, a social reflection common to Wikipedias, and with connections between Homer and Dante, a figure more central to Italian literary identity. The English page reflects a concern with the Odyssean journey and its possible real-world correspondences. More commonality was found in another example comparing the discussion pages of J.D Salinger's *The Catcher in the Rye*. These alignments and misalignments of reader's concerns can destabilize the primacy of concerns held by a given reading community and speaks to one of the core benefits of reading world literature.

Bibliography

Bayliss, G. (2013). Exploring the Cautionary Attitude toward Wikipedia in Higher Education: Implications for Higher Education Institutions. *New Review of Academic Librarianship*, **19**(1): 36-57.

Callahan, E. and Herring, S. (2011). Cultural Bias in Wikipedia Content on Famous Persons. *Journal for Information Science and Technology*, **62**(10): 1899-1915.

Ensslin, A. (2011). 'What an un-wiki way of doing things': Wikipedia's Multilingual Policy and Metalinguistic Practice. *Journal of Language & Politics*, **10**(4): 535-61.

Filatova, E. (2012). Information Overlap in Multilingual Wikipedia and Summarization. *International Journal of Cooperative Information Systems*, **21**(4): 383-403.

Goldberg, D., Looney, D. and Lusin, N. (2015). Enrollments in Languages Other than English in United States Institutions of Higher Education, Fall 2013. *Modern Language Association*. First published online February 2015. https://apps.mla.org/pdf/2013_enrollment_survey.pdf.

Gorard, S. and Selwyn, N. (2015). Students Use of Wikipedia as an Academic Resource—Patterns of Use and Perceptions of Usefulness. *The Internet and Higher Education* , **28**: 28-34.

Kashima, Y. (2000). Maintaining Cultural Stereotypes in the Serial Reproduction of Narratives. *Personality and Social Psychology Bulletin*, **26**(5): 594-604.

Knight, C. and Pyke, S. (2012). Wikipedia and the University, a case study. *Teaching in Higher Education*, **17**(6): 649-59.

Li, H., Graesser, A. C. and Cai, Z. (2014). Comparison of Google translation with human translation. *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference*. Pensacola Beach, FL, May 2014, pp. 190-5.

Ni, X., Sun, J., Hu, J. and Chen, Z. (2009). Mining multilingual topics from wikipedia. *Proceedings of the 18th international conference on World Wide Web*. Madrid, Spain, April 2009.

Nothman, J., Ringland, N., Radford, W., Murphy, T. and Curran, J. (2013). Learning Multilingual Named Entity Recognition from Wikipedia. *Artificial Intelligence* , **194**: 151-75.

Purdy, J. (2009). When the Tenets of Composition Go Public: A Study of Writing in Wikipedia. *College Composition and Communication*, **61**(2): W351-W373.

Sweeney, M. (2012). The Wikipedia Project: Changing Students from Consumers to Producers. *Teaching English in the Two-Year College*, **39**(3): 256-67.

Venuti, L. (1998). *Scandals of Translation: Towards an Ethics of Difference*. London: Routledge.

Vetter, M. (2014). Archive 2.0: What Composition Students and Academic Libraries Can Gain from Digital Collaborative Pedagogies. *Composition Studies*, **42**(1): 35-53.

Exploring and Discovering Archive-It Collections with Warcbase

Ian Milligan
i2milligan@uwaterloo.ca
Department of History, University of Waterloo, Canada

Jimmy Lin
jimmylin@uwaterloo.ca
David R. Cheriton School of Computer Science, University of Waterloo, Canada

Jeremy Wiebe
jrwiebe@uwaterloo.ca
Department of History, University of Waterloo, Canada

Alice Zhou
alice.zhou@uwaterloo.ca
David R. Cheriton School of Computer Science, University of Waterloo, Canada

Introduction

Big Data is reshaping the historical profession in ways we are only now beginning to grasp. The growth of digital sources since the advent of the World Wide Web in 1990-91 presents new opportunities for social and cultural historians. Large web archives contain billions of webpages, from personal homepages to professional or academic websites, and now make it possible for us to develop large-scale reconstructions of the recent web. Yet the sheer number of these sources presents significant challenges: if the norm until the digital era was to have human information vanish, "now expectations have inverted. Everything may be recorded and preserved, at least potentially" (Gleick, 2012).

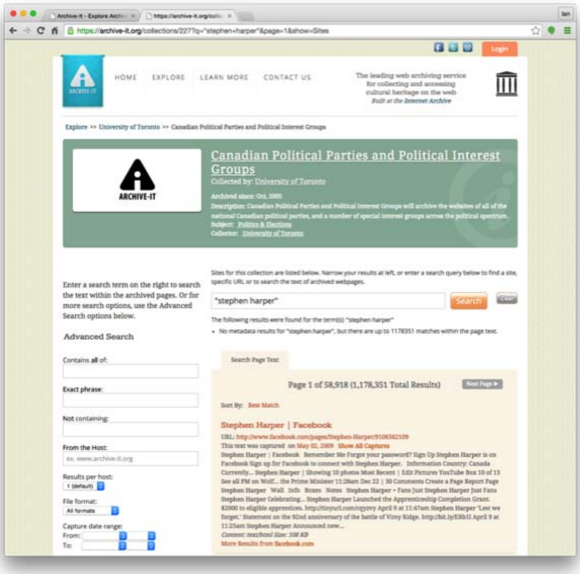
While the Internet Archive makes archived web content available to the general public and mainstream scholarly community through its "Wayback Machine," (at http://archive.org/web) which allows visitors to enter a Uniform Resource Locator (URL) to visit archived web versions of a particular page, this system is limited: not only do visitors need to know the URL in the first place, but they are limited to individual readings of single webpages.

By unlocking the Wayback Machine's underlying system of specialized files, primarily ISO-standardized WebARChive (WARC) files, we can develop new ways to systematically track, visualize, and analyze change occurring over time within web archives. Warcbase, an open-source platform for managing web archives built on Hadoop and HBase, provides a flexible data model for storing and managing raw content as well as metadata and extracted knowledge. Tight integration with Hadoop provides powerful tools for analytics and data processing. Using a case study of one collection, this paper introduces the work that we have been doing to facilitate web archive access with warcbase. We have growing documentation at http://docs.warcbase.org.

Project Rationale and Case Study

In 1996, the Internet Archive launched a complementary research services company, Archive-It, which offers subscription-based web archiving to collecting institutions.

The University of Toronto Library (UTL) began collecting a quarterly crawl in 2005 of Canadian political parties and political interest groups (the collections were separate in 2005, merging in 2006) (University of Toronto, 2015). The collection itself has a murky history: UTL had been part of a broader project that would have collected political websites. It fell through, but UTL opted to carry out their crawl on their own and the librarian was responsible for selecting the seed list herself (faculty and other librarians did not respond for calls for engagement). While formal political parties are robustly covered, the "political interest groups" collection was a bit more nebulous: sites were discovered through keyword searches, and some were excluded due to robots.txt exclusion requests. Beyond this brief sketch, we have little information about the decisions made in 2005 to create this collection. This lack of documentation is a shortcoming of this collection model, as if a historian was to use this material in a peer-reviewed paper, questions would be raised about its representativeness.


Archive-It - Explore Archi
https://archive-it.org/colle
https://archive-it.org/collections/227?q="stephen+harper"&page=1&show=Sites
Login
HOME
EXPLORE
LEARN MORE
CONTACT US
ARCHIVE-IT
The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive
Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups
ARCHIVE-IT
Search
Clear
Search Page Text
Next Page ▶

Canadian Political Parties and Political Interest Groups
Collected by: University of Toronto
Archived since: Oct, 2005
Description: Canadian Political Parties and Political Interest Groups will archive the websites of all of the national Canadian political parties, and a number of special interest groups across the political spectrum.
Subject: Politics & Elections
Collector: University of Toronto

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains all of:

Exact phrase:

Not containing:

From the Host: ex. www.archive-it.org

Results per host: 1 (default)

File format: All formats

Capture date range:
From:
To:

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

"stephen harper"

The following results were found for the term(s): "stephen harper"

- No metadata results for "stephen harper", but there are up to 1178351 matches within the page text.

Page 1 of 58,918 (1,178,351 Total Results)

Sort By: Best Match

Stephen Harper | Facebook
URL: http://www.facebook.com/pages/Stephen-Harper/9106562109
This text was captured on May 02, 2009 Show All Captures
Stephen Harper | Facebook Remember Me Forgot your password? Sign Up Stephen Harper is on Facebook Sign up for Facebook to connect with Stephen Harper. Information Country: Canada Currently... Stephen Harper | Showing 10 photos Most Recent | Edit Pictures YouTube Box 10 of 13 See all PM on Wolf... the Prime Minister 11:28am Dec 22 | 30 Comments Create a Page Report Page Stephen Harper Wall Info Boxes Notes Stephen Harper + Fans Just Stephen Harper Just Fans Stephen Harper Celebrating... Stephen Harper Launched the Apprenticeship Completion Grant. $2000 to eligible apprentices. http://tinyurl.com/ogyzvy April 9 at 11:47am Stephen Harper 'Lest we forget.' Statement on the 92nd anniversary of the battle of Vimy Ridge. http://bit.ly/E8b1l April 9 at 11:25am Stephen Harper Announced new...
Content: text/html Size: 108 KB
More Results from facebook.com

Figure 1: Archive-It Search Portal

If a user wants to use the Canadian Political Parties and Interest Groups Collection (CPP) through Archive-It today, they visit the collection page at https://archive-it.org/collections/227 and enter full-text search queries. In August 2015, our group also launched http://webarchives.ca, based on the British Library's SHINE front end for web archives; this was a way to facilitate a different form of more casual user access, aimed at the general public (we discuss this in a separate paper).

The Archive-It portal is limited. There are no readily-available metrics of how many pages have been collected, how they break down by domain and date, and the portal undoubtedly provides skewed results unless the search phrase is dramatically narrowed down.

Consider the search for "Stephen Harper," Canada's Prime Minister between 2006 and 2015 in Figure 1.The results are decent: Harper's Facebook page from 2009, a Twitter snapshot from 2010, and some long-form journalism articles and opposition press releases. But amidst the 1,178,351 results, there is no indication as to how the ranking took place, what facets are available, and how things may have changed over the last ten years of the crawl.

The data is there, but the problem is access.

Warcbase: A Platform for Web Archive Analysis

Warcbase is a web archive platform, not a single program. Its capabilities comprise two main categories:

1. Analysis of web archives using the Pig or Spark programming languages, and assorted helper scripts and utilities

2. Web archive database management, with support for the HBase distributed data store, and OpenWayback integration providing a friendly web interface to view stored websites

One can take advantage of the analysis tools (1) without bothering with the database management aspect of Warcbase – in fact, most digital humanities researchers will probably find the former more useful. This paper focuses on the former capabilities, showing how we can use the warcbase platform to carry out text and network analyses.

Using Warcbase on Web Archival Collections: Text Analysis

We have begun to document all warcbase commands on a GitHub wiki, found at https://github.com/lintool/warcbase/wiki. We begin with installation instructions, and then provide simple scripts written in Apache Spark to run the commands.

While possible to generate a plain text version of the entire collection, a more fruitful approach has been to generate date-ordered text for particular domains. If a researcher is interested in say, the Green Party of Canada's evolution between 2005 and 2015, they can extract the plain text for greenparty.ca by running the following script:

```
import org.warcbase.spark.matchbox.{RemoveHTML, RecordLoader}
import org.warcbase.spark.rdd.RecordRDD._

RecordLoader.loadArc("src/test/resources/arc/example.arc.gz", sc)
  .keepValidPages()
  .keepDomains(Set("greenparty.ca"))
  .map(r => (r.getCrawldate, r.getDomain, r.getUrl, RemoveHTML(r.
  getContentString)))
  .saveAsTextFile("out/")
```

All they would need to change would be the path/to/input to the directory with their web archive files, the path/to/output for where they want to save the resulting plain-text files, and the greenparty.ca value to whatever domain they might be interested in researching.

They then receive a date-ordered output of all plain text for that domain (as per the extractCrawldateDomainUrlBody command). It can then be sorted and used in other research avenues. For example, this plain text could be loaded into a text analysis suite such as http://voyant-tools.org/ or other digital humanities environments.

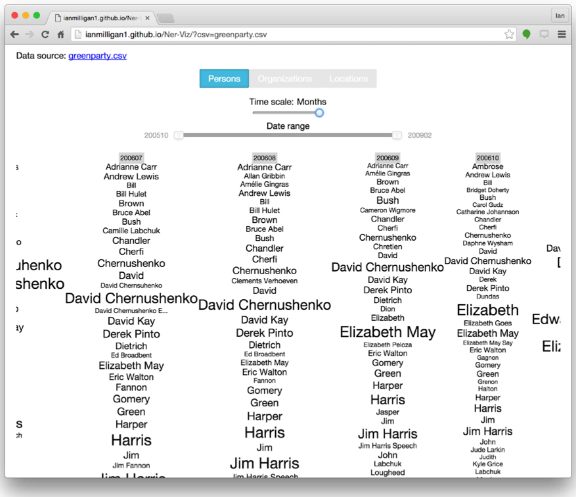


Figure 2: Named Entity Visualization within Warcbase

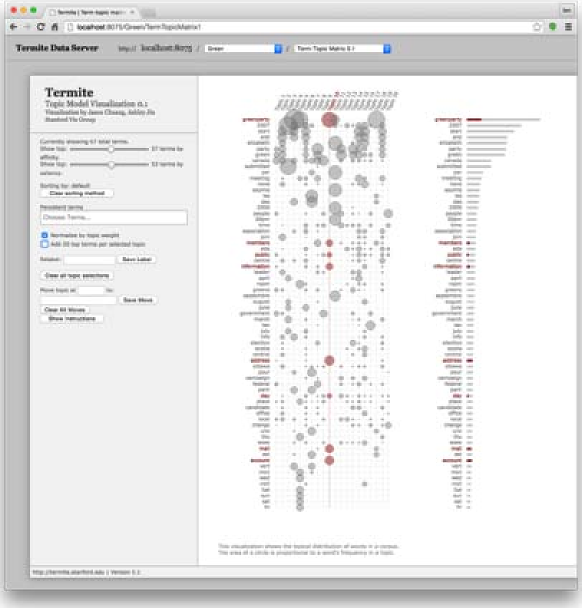

Termite | Term-topic matrix
localhost:8075/Green/TermTopicMatrix1
Termite Data Server
localhost:8075
Green
Term Topic Matrix 0.1
Termite
Topic Model Visualization

Figure 3: Termite Topic Model

We have also been experimenting with other visualizations based on the extracted plain text. Computationally intensive textual analysis can be carried out using warcbase itself. Using the Stanford NER package in parallel, we have a script that extracts entities, counts them, and then visualizes them using D3.js to help see overall changes in a web archival collection. Figure 2 below shows the output of the NER visualizer.

Finally, another text approach is topic modelling (Blei et al., 2003). LDA works by finding topics in unstructured text. To visualize topic models, we elected to use the Termite Data Server, which is a visual analysis tool for exploring the output of statistical topic models ("uwdata/termite-data-server," n.d.). As Figure 3 demonstrates, the visualization allows you to get a top-down view at the topics found in a web archive.

Warcbase presents versatile opportunities to extract plain text and move it into other environments for analysis. Unlike the keyword-based Archive-It portal, we now have data that can be inquired in many fruitful ways.

Using Warcbase on Web Archival Collections: Hyperlink Analysis

Warcbase can also extract hyperlinks. While text can be very important, these sorts of metadata can often be more important: allowing us to see changes in how groups link to each other, what articles and issues were important, and how relationships changed over time.

Consider Figure 4, which visualizes the links stemming from and between the websites of Canada's three main political parties.

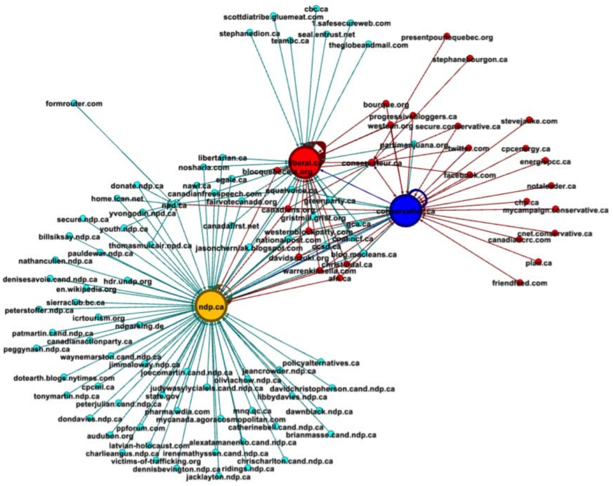


Figure 4: Three major political parties in Canada

Above, we can see which pages only link to the left-leaning New Democratic Party (ndp.ca), those that link only to the centrist Liberals (liberal.ca) in the top, and those that only connect to and from the right-wing Conservative Party at right. We can use it to find further information, such as in Figure 5.

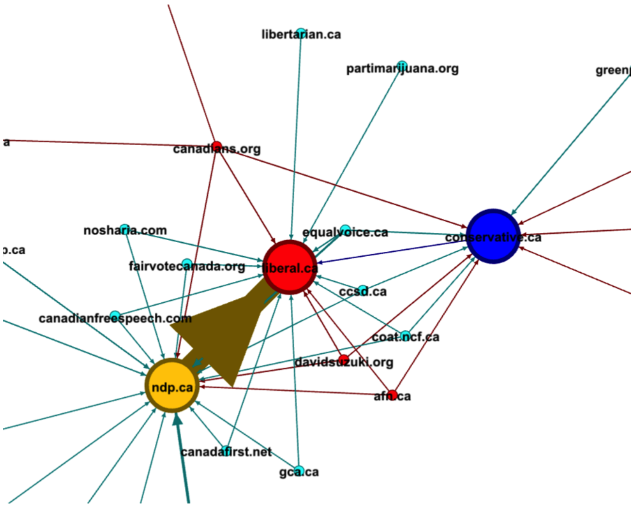
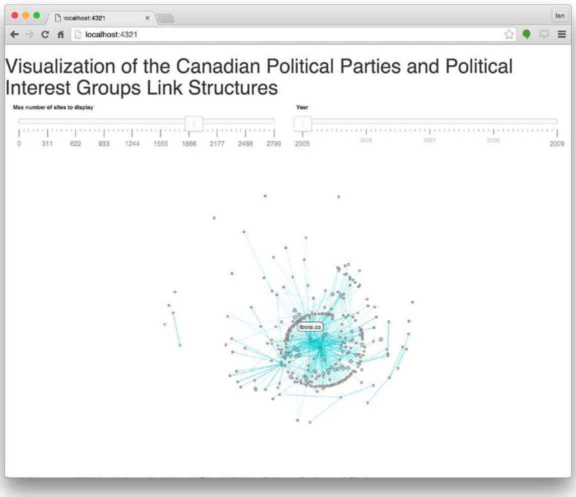


Figure 5: NDP attack

The above links are from the 2006 Canadian federal election. The Liberal Party was then in power and was under attack by both the opposition parties. In particular, the left-leaning NDP linked hundreds of times to their ideologically close cousins, the centrist Liberals, as part of their electoral attacks, ignoring the right-leaning Conservative Party in the process. Link metadata illuminates more than a close reading of an individual website would. It contextualizes and tells stories itself.



Visualization of the Canadian Political Parties and Political Interest Groups Link Structures

Figure 6: Link Visualization

While we have traditionally used Gephi to do analysis, importing material into Gephi from warcbase required many manual steps as documented at https://github.com/lintool/warcbase/wiki/Gephi:-Converting-Site-Link-Structure-into-Dynamic-Visualization. We have been prototyping a link analysis visualization in D3.js, which can run in browser (Figure 6).

Conclusions

With the increasingly widespread availability of large web archives, historians and Internet scholars are now in a position to find new ways to track, explore, and visualize changes that have taken place within the first two decades of the Web. Warcbase will allow them to do so. This project is among the first attempts to harness data in ways that will enable present and future historians to usefully access, interpret, and curate the masses of born-digital primary sources that document our recent past.

Bibliography

Brügger, N. (2008). The Archived Website and Website Philology: A New Type of Historical Document, *Nordicom Review*, **29**(2): 155–75.

Gleick, J. (2012). *The Information: A History, a Theory, a Flood*. London: Vintage.

University of Toronto. (2015). Archive-It - Canadian Political Parties and Political Interest Groups [WWW Document]. https://archive-it.org/collections/227.

uwdata/termite-data-server. GitHub. https://github.com/uwdata/termite-data-server.

Blei, D. M., Ng, A. Y., Jordan, Michael I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res*, **3**: 993–1022.

Brügger, N. and Finnemann, N. O. (2013). The Web and Digital Humanities: Theoretical and Methodological Concerns. *Journal of Broadcasting & Electronic Media*, **57**(1): 66–80.

Lin, J., Gholami, M. and Rao, J. (2014). Infrastructure for Supporting Exploration and Discovery in Web Archives. *Proceedings of the 23rd International Conference on World Wide Web*. doi:10.1145/2567948.2579045.

Representación De Otras Literaturas Mexicanas En Medios Digitales

Ernesto Miranda
mirandatrigueros@gmail.com
INAH, Mexico

Introducción

Este trabajo tiene como objetivo mostrar las posibilidades que ofrecen las herramientas digitales para representar literaturas de México pertenecientes a los pueblos originarios. Estas formas literarias han sido subrepresentadas y los acercamientos a ellas siempre han estado marcados por una visión parcial sobre los procesos intrínsecos que las caracterizan y que las distinguen de las literaturas oficiales de México.

Las conclusiones se presentan en base a los resultados de dos proyectos: por un lado la edición digital *Códice Mendoza* y por otro un prototipo digital para la representación de poesía ritual. Ambos proyectos, como se verá, ilustran las ventajas de la representación digital sobre la que se ha venido haciendo por décadas en soportes impresos.

Antecedentes

México es un país con una profunda tradición intelectual que viene desde mucho tiempo atrás de la llegada de los españoles. Las culturas que habitaron lo que hoy se denomina Mesoamérica lograron un enorme desarrollo científico y cultural, que se ve reflejado en su organización política, en su esplendor arquitectónico, en la precisión de sus cálculos astronómicos y en sus expresiones artísticas y literarias.

Es importante mencionar que hoy en día ese conocimiento sigue vivo a través de más de 60 culturas que cohabitan en el territorio mexicano. Esas culturas, en mayor o menor medida, son portadoras de ese conocimiento milenario, y siguen representando y actualizando ese saber a través de sus literaturas.

El investigador británico Gordon Brotherston se ha referido a estas literaturas como los libros del "cuarto mundo", expresiones literarias que no necesariamente responden a los estándares que hemos definido durante siglos en la tradición intelectual hegemónica como literatura. (Brotherston, 1997)

Entre los ejemplos que podemos encontrar a lo largo del continente se encuentran los quipús andinos, y por supuesto los dos casos que nos ocupan: la poesía ritual y los manuscritos pictóricos mesoamericanos. Vale la pena recordar que muchas de estas culturas tenían una forma integral y holística de representar el cosmos, por lo que estas expresiones no se entienden de manera aislada de otros factores culturales como los rituales, la guerra, la caza o los fenómenos naturales.

Estos ejemplos de formas literarias no canónigas, que se multiplican por todo el continente americano, deben ser reconocidos como textos en sí mismos. Pueden ser entendidos como una modalidad enmarcada de lenguaje y todos cuentan con coherencia interna (estructura, formato y orden); codificación que le permite resistir al paso del tiempo; todas son entidades en sí mismas y cuentan con una estructura y una función comunicativa.

Es muy probable que los mecanismos de representación que hemos usado hasta ahora hayan sido poco sensibles a las necesidades formales que este tipo de expresiones exigen. Debemos reconocer que nuestros acercamientos están siempre enfocados a través de una visión binaria entre lo oral y lo escrito, irrelevante e insuficiente para las literaturas de la América Indígena.

En ese sentido hemos sido indiferentes o insensibles al ignorar todas estas categorías de representación. Si bien las hemos estudiado, y gracias a una enorme tradición de estudiosos desde fray Bernardino de Sahagún, pasando por el Padre Ángel María Garibay hasta llegar a Miguel León Portilla y Alfredo López Austin, las pudimos conocer, no hemos resuelto aún como adecuarlas para que su representación formal responda a la complejidad de su esencia.

Es necesario que entendamos que estos textos mesoaméricanos son mucho más complejos de lo que hoy leemos como medios escritos. Son indisociables de todos los elementos que lo integran: material, oral y ritual. Son diacrónicos y dialógicos y al mismo tiempo multimediáticos.

Por último es necesario tener en cuenta que las dos formas expresivas de las que nos ocupamos, tienen origen en el México prehispánico, y en algunos casos han sobrevivido a los procesos de aculturación. También, estas dos expresiones resultan únicas por sus características y, hasta el momento en que estas líneas se escriben, el autor no ha podido conocer otros proyectos que aborden desde una perspectiva holística el problema de representación de estas literaturas o expresiones culturales mexicanas en otro tipo de medios.

Representación de códices mexicanos en medios digitales

El primer proyecto que se abordará es la edición digital del *Códice Mendoza* esfuerzo histórico por ofrecer en plataformas digitales un nuevo acercamiento a estos manuscritos históricos. Estos documentos pictográficos son considerados objetos culturales complejos que representaban el conocimiento histórico, simbólico, calendárico, económico y ritual de los antiguos mexicanos. Fueron creados en diferentes tipos de soportes como el papel amate, el papel de maguey o el papel europeo, después de la llegada de los españoles.

La escritura de los códices mexicanos no es fonética salvo en el caso de los códices mayas. En el caso de los códices zapotecos y mexicas la relación entre las imágenes es lo que determina el significado. Este tipo de escritura no alfabética ha sido definida como semasiográfica (Boone Hill, 1994, 20). Asimismo, integran en un todo holístico lo que nosotros podríamos considerar como elementos separados. En ese sentido, los podemos considerar como textos multimedia ya que echan mano de imágenes, oralidad, memoria, ritual e incluso elementos no directa o físicamente asociados como esculturas, sitios sagrados o algún corpus de inscripciones. De igual manera, la lectura de estos documentos era un acto performático y ritual, sólo ejecutado por quien Jerome Rothenberg llamaría un técnico de lo sagrado.

En síntesis podemos decir que los códices son soportes abiertos, dinámicos, participativos, interactivos y 'suceden' en tiempo real. Por estas características resulta muy difícil imaginar que podamos representarlos en un soporte plano bidimensional como una edición en papel.

El especialista Miguel León-Portilla ha considerado que es insuficiente traducir un códice a una edición en papel. Quizá lo mismo se puede decir de cualquier traducción y transliteración, aunque por las características descritas arriba, se acentúa la complejidad en el caso de los códices. León-Portilla también estableció una interesante analogía, muy adecuada para nuestro tema, donde compara a los CR-ROMs con los códices. Enfatiza que ambos tienen varias formas de lectura, los dos son multimedia al combinar imágenes, texto y sonido, al tiempo que ambos tiene un afán totalizados de diversas facetas del conocimiento. (León-Portilla, 2003)

Hoy, el equivalente a un CD-ROM sería una aplicación web o una aplicación nativa para algún dispositivo móvil.

En este sentido, la edición digital del *Códice Mendoza* demostró que la representación digital es óptima para los códices mexicanos. Su complejidad semántica y también su carácter multimedial se representan de manera más holística en medios digitales. En este sentido vale la pena recordar las palabras de Willard McCarty cuando señala que toda traducción de un medio textual a un medio digital, conlleva una importante pérdida de sentido (McCarty, 2008). Sin embargo, en este caso la pérdida es mucho menor ya que lo que estamos trasladando es un medio dinámico a otro medio dinámico.

Del mismo modo, consideramos que el resultado funciona de manera doble en varios sentidos: permite un acercamiento al público en general y también un acercamiento a los académicos especializados. Al mismo tiempo, permite una representación más completa pero también las herramientas tecnológicas nos permiten un estudio más profundo de estos documentos, principalmente por su accesibilidad.

Representación de poesía ritual en medios digitales

Dentro del territorio mexicano existen más de sesenta lenguas indígenas, que con sus variaciones dialectales son más de trescientas. Como en todas las lenguas del mundo, son complejísimas las expresiones literarias verbales que residen dentro de estas lenguas y estas culturas.

Estas expresiones son consideradas como patrimonio intangible, inmaterial o vivo y pese a que existe una buena cantidad de registros, son escasos los que recaen en su valor literario, y en muchas ocasiones se desdeña este tipo de acercamientos frente a los de las ciencias sociales.

Al igual que los códices, estas expresiones son dinámicas, abiertas, multimodales, participativas e interactivas. Dentro de la enorme diversidad de expresiones, se encuentran los cantos chamánicos o rituales, que por sus características son aún más complejos. En ellos, también inciden aspectos rituales, estados de trance y otros elementos contextuales difíciles de representar como pintura facial, eructos o incluso convulsiones.

Hasta hace poco tiempo estas expresiones se representaban y se estudiaban a través de ediciones impresas. En palabras del teórico John Miles Foley, esta forma de trabajar con las formas verbales aplana o diseca su expresividad. Asimismo, los acercamientos académicos han mostrado poco interés por las características poéticas de estas expresiones[1]. Fue gracias a enfoques novedosos y propositivos como la etnopoética, que se logró un cambio de actitud hacia cómo trabajar estos materiales. (Miranda Trigueros, 2007).

Al igual que en el caso de los códices, los medios digitales parecen mucho más adecuados para representar estas expresiones orales. Podemos ser más fieles en la representación, incluir un mayor número de canales de significado, mayor contexto y ser más transparentes con la enunciación original.

En respuesta a estas necesidades el que escribe estas líneas creó el prototipo DARP (Digital Archive of Ritual Poetry) como tesis de maestría en 2013, para obtener el grado de maestro en humanidades digitales en el King's College de Londres. . El prototipo está basado en un fragmento del canto chamánico de la reconocida sanadora mazateca María Sabina.

El resultado que nos ofreció el prototipo es un medio digital que nos permite representar de una mejor manera este tipo de expresiones. A diferencia de esfuerzos anteriores, donde, por ejemplo, se representaban en papel acompañados de un CD de audio (Miranda Trigueros, 2007), la integración total de la representación aural con el texto lo hace que funcione mucho mejor. El uso de una aplicación simple basada en Javascript y JQuery permitió sincronizar el audio al texto en una lectura interactiva. Asimismo se integraron otros elementos hipermediales que amplían la comprensión del texto y su significado.

Por otro lado, la creación del prototipo permitió comprobar que la codificación del texto permite un acercamiento totalmente novedoso al contenido, obligando al investigador a hacerse preguntas hasta ese momento inéditas (Sperberg-McQueen, 2009, 35). La codificación con TEI y XLST permitió la creación de índices automatizados que estructuran el contenido textual y ofrecen herramientas al investigador para hacer un análisis mucho más ágil y certero que si lo pensáramos en términos manuales.

El prototipo busca convertirse en un recurso más sólido para lo cual se plantean dar los siguientes pasos:

- definir una ontología básica de los elementos que pueden encontrarse dentro de este tipo de expresiones.
- Integrar un corpus de diferentes expresiones que se encuentren a lo largo del territorio nacional
- Creación de una base de datos consultable para poder encontrar elementos comunes y desde allí establecer una poética de los cantos rituales mexicanos.

Conclusiones

Las herramientas digitales nos permiten representar de mejor manera los llamados libros del cuarto mundo.

Podemos crear ediciones más holísticas, precisas, fieles y con un mayor grado de contexto que permite una mejor comprensión del objeto representado.

Se puede decir también que después de haber trabajado en dos casos disímiles en su formato, pero análogos en su origen y en su complejidad, la representación de un medio dinámico siempre debe de ser representada en un medio dinámico.

Tanto en el caso de la poesía oral ritual como en el caso de los códices se hace patente que la complejidad de las expresiones culturales prehispánicas encuentra una representación más rica y transparente a través de las herramientas digitales. Sus características dinámicas, abiertas, procesuales, interactivas y multimediales obligan a pensar en los medios digitales como óptimos para la representación y estudio de este tipo de literaturas. Los recursos digitales son más fieles al original, abren más canales semánticos, ofrecen más contexto, y una forma más holística y transparente de presentación.

Asimismo, en ambos casos encontramos enormes posibilidades para los interesados en el estudio de la cultura mexicana, para entender mejor este enorme patrimonio, esta literatura sumergida, donde creemos que es necesario profundizar en las ventajas y virtudes que nos ofrecen las herramientas digitales para un replanteamiento total en el estudio de ellas.

Bibliography

Brotherston, G. (1997). *La América indígena en su literatura: los libros del cuarto mundo.* México, FCE.

Boone Hill, E. and Mignolo, W. (eds). (1994). *Writing Without Words: Alternative Literacies Mesoamerica & the Andes*. EUA, Duke University Press.

Foley, J. M. (2012). *Oral Tradition and the internet. Pathways of the mind*. EUA, University of Illinois Press.

León-Portilla, M. (2003). *Códices, los antiguos libros del nuevo mundo.* México, Aguilar.

Miranda Trigueros, E. (2007). *La etonopoética y los cantos de María Sabina, una aproximación,* (tesis de licenciatura). México, UNAM.

McCarty, W. (2008). What's going on?. *Literary and Linguistic Computing*, **23**(3): 253-61.

Rothebnberg, J. and Tedlock, D. (1970). Statement of Intention. *Alcheringa*, **1**: 5.

Sperberg-McQueen, C. M. (2009). How to teach your edition how to swim *Literary and Linguistic Computing*, **24**(1): 27-39.

Notes

[1] Hay varios factores que inciden en la falta de preocupación por representar de manera más completa estas expresiones, principalmente hay poco interés de parte de los antropólogos por las características meramente poéticas de este tipo de literaturas En general, las consideran como un accesorio para conocer otros aspectos de la cultura a la cual pertenecen, una fuente más para el estudio del otro. (Rothenberg and Tedlock, 1970).

Toward A Use-Value Paradigm For The Sustainability Of Digital Research

Francesca Morselli
francesca.morselli@gmail.com
Data Archiving and Networked Services (DANS), Den Haag, The Netherlands

Jennifer Edmond
EDMONDJ@tcd.ie
Trinity College Dublin, Ireland

What is Sustainability?

Sustaining the results of digital humanities research projects remains an ongoing challenge within the wider DH ecosystem. This is particularly the case for research infrastructure developments, where the scale, complexity and indeed the overarching aim of the project to serve a potentially still emerging research community makes their continued accessibility even more important, and even more difficult.

Like any enduring challenge, sustainability of complex digital resources has been subjected to a certain amount of scholarly investigation (though less overall than one might expect). What a survey of this literature shows, however, is that the perspectives on how to face the challenge of sustainability are highly dependent on how a project or infrastructure views itself: for example, when viewed primarily as a *form of organisation or institution*, the sustainability model proposed will likely focus on the necessary 'business model' for maintaining the services created (Maron et al., 2009). The Archives Portal Europe (APE[1]), for example, established itself as a Foundation after the end of the project, in order to maintain the vital functions of the infrastructure and to further connect with other projects and similar initiatives.

Alternatively, when project outputs are viewed as a *tool or technical platform*, a sustainability proposal will primarily take into consideration issues and practices such as migration and curation of elements such as the repositories where the data are stored and continued maintenance of work environments and specific tools. The TextGrid project[2], for example, has been hugely successful in rolling its activities forward over a long period, continuing to make its services available to users. At its best, this approach results in a broad focus on software durability, documentation of processes and the modularity of services (Buddenbohm et al., 2015).

But to understand sustainability thoroughly we must also engage a second huge and unresolved issue in digital humanities, that is the reuse of project outputs and data. In particular, the "Log Analysis of Digital Resources in the

Arts and Humanities, or the LAIRAH project (Warwick et al., 2008) has contributed significantly to our understanding of what factors enable digital projects and tools to be found and adopted by users. From the results of this project we can see another model for the sustainable project to emerge, in which the *communication and branding* of the project is a key element of its success.

The CENDARI Project and its Approach to Sustainability

This presentation brings forward the hypothesis that a successful approach to sustainability for Research Infrastructures needs to be comprehensive; an approach that doesn't just consider data or technology, community, communications or processes, but in fact all of them simultaneously. In addition, it should focus not only on a project as a collection of tangible and intangible assets, but also on the potential user base for these assets, and what these users consider valuable about them.

Discussion of this user-centred approach to sustainability will be based on the experiences of the Collaborative Digital Archival Research Infrastructure (CENDARI[3]) project's year-long sustainability planning exercise, conducted from January 2014-January 2015. This exercise, which built upon previous work in the project and a strong link to the Digital Research Infrastructure for the Arts and Humanities (DARIAH ERIC[4]), resulted in a set of principles and processes for mapping and sustaining user value from the project for the medium and long terms. Although both the generic process (which will be released as a sustainability toolkit at the end of the project) and the specific actions implemented by the project match on some level the specifics of the CENDARI development, they also reflect the reality, identified by Joris Van Zundert, of the "fluidity" of research infrastructure, caught up in both the digital information lifecycle and the creation of knowledge by end users, as well as the software components (Van Zundert, 2012).

The Sustainability Planning Cycle

The CENDARI sustainability planning process was comprised of a series of 4 stages, from pre-planning to closure and post-project actions, each of which contributed to the overall, holistic sustainability strategy. This cycle was intended to counteract a natural impetus within projects to view sustainability as a concern only for the final phase of the project, rather than one to be integrated into the project's development and even its conception.

The CENDARI Assets and Multilevel Sustainability Action Plan

As a key component of the second and third phases of the CENDARI project sustainability planning process, the project carried out a thorough audit (including a stakeholder validation meeting) to refine its understanding of what assets the project had generated and how they could be maintained, shared and indeed passed on to its key users for further development. This process identified 7 categories of assets as most likely to find future usage, each of which posed unique challenges in how they could be captured, made visible and sustained. It has been one of the greatest challenges of the CENDARI sustainability planning process to ensure that for each of these areas we could find a solution, as we would for our personal work data, to make them findable and reusable in a contextualised manner, and preserve them in 'multiple formats and multiple locations.' For each asset type, the audience for potential future use is different, and therefore the solution proposed is as well.

The CENDARI **portal** is the most visible of its assets, representing the final synthesis of the project's activities and its main point of access. For many projects, this would be where sustainability planning would not only begin, but end. CENDARI approached this sustainability challenge via a three pronged strategy, guaranteeing 3 years of access through the German arm of DARIAH but also ensuring new communities and new approaches would be recruited to continue development.

But the portal is not only useful in its complete final form, but also as a collection of unique **services, tools and components** optimised to support DH research. This possible reuse of the project outcomes was foreseen from the beginning, and a very modular, service oriented architecture was adopted for the project. The tools therefore require a sustainable pathway outside of the portal. That said, however, connecting tools with potential user bases is a constant challenge. The software community practice of using GitHub to share software was adopted, but further awareness raising was also required to ensure the maximal future use for the tools.

CENDARI holds a lot of **data** from different sources, some unique to the project, others well signposted elsewhere, and with different requirements and expectations for sustainability. This has been its legacy as a project seeking to reuse archival data for historical research, where the culture and ability to share data is unevenly developed. The CENDARI data portal gives access to this data, and the project's data agreement and license have been developed with DARIAH as a co-signatory, so in many ways DARIAH had already agreed from an early point in the project to sustain this data. But DARIAH is not well-known as a data provider or source, and this solution alone may not maximise visibility and reuse. Therefore a redeposit protocol for unique data with an external trusted source has also been facilitated.

The **Archival Research Guides** exist as a particular subset of the data unique to CENDARI, but their status as both primary and secondary research sources justifies their consideration as an asset class in themselves, in particular because of the manner in which they challenge existing norms of publication, communication and evaluation in the discipline of history. As extended and enhanced publications, incorporating analysis, links to data sources, multimedia objects, and links to project ontologies, these guides need to be delivered within the project portal. But to sustain these unique works of scholarship only in that format would again potentially limit their visibility. They will therefore be offered in one or more export formats, as well as becoming the focus of both a review publication and a research paper to be submitted to a mainstream (not digital) historical journal. In this way, their contributions to scholarship can be recognised as independent from the format in which they have been delivered.

Given how particular many of the project experiences in building for the DH community had been, a specific audit of CENDARI's **tacit knowledge** was also undertaken, and several white papers and process oriented toolkits have emerged from the project on the foot of this (including, for example, a 'White Book of Archives' documenting the project's experience of federating highly heterogenous data from traditional collection holding institutions). As a related issue, some of the project's **management assets** may also have a future utility for others.

Perhaps the least easily defined and sustained aspects of the CENDARI project will be the **communities** - mixed and homogenous groups of historians and other humanistic scholars, collections experts and technologists - it has brought together and formed. Interconnectivity between cognate projects will be a key resource for this, as some communities will have interests across these projects, and networks can and should be shared. But some of the community aspects are very unique to CENDARI, and will have a specific role in guiding the future use of the portal and its components: for this reason, the project will use another DARIAH mechanism, the working group, to provide a structure for continued development of project concerns and assets.

As can be seen from this description, CENDARI has made both its own sustainability and the potential future role of the DARIAH ERIC in the sustainability of medium-to large-scale digital projects in Europe into key areas of applied research and development. The resulting tool-kit for sustainability will hopefully assist future projects in extending both their sustainability planning and strategies in the future.

Bibliography

Buddenbohm, S., Enke, H., Hofmann, M., Klar J., Neuroth H. and Schwiegelshohn, U. (2015). Success Criteria for the Development and Sustainable Operation of Virtual Research Environments, *D-Lib Magazine*, **21** 9/10. doi:10.1045/september2015-buddenbohm.

Dempsey, L.. (2000). The Subject Gateway: Experiences and Issues Based on the Emergence of the Resource Discovery Network, *Online Information Review*, **24**(1): 8–23.

European Science Foundation. (2011). Research Infrastructure in the Digital Humanities. http://www.esf.org/index.php?eID=tx_nawsecuredl&u=0&g=0&t=1456595004&hash=734ba001ffd70284abd6390009ca0842e16fab98&file=fileadmin/be_user/research_areas/HUM/Strategic_activities/RIs_in_the_Humanities/SPB42_44p-5oct_FINAL.pdf (accessed August 2015).

Geyer-Schulz, A., Neumann A., Heitmann A. and Stroborn, K. (2006). Strategic Positioning Options for Scientific Libraries in Markets of Scientific and Technical Information - the Economic Impact of Digitization, *Journal of Digital Information*, **4**: 2. https://journals.tdl.org/jodi/index.php/jodi/article/view/101 (accessed September 2015).

Giarlo, M. (2013). Academic Libraries as Data Quality Hubs, *Journal of Librarianship and Scholarly Communication*, **1**: 3. doi:http://dx.doi.org/10.7710/2162-3309.1059. (accessed September 2015).

JISC Digital Media. Sustainability of Digital Collections: A Practical Guide. http://www.jiscdigitalmedia.ac.uk/infokit/digitisation-funding-and-sustaina/sustainability-of-digital-collections-a-practical-guide (accessed October 2015).

Maron N. L., Smith K. K. and Loy, M. (2009). Sustaining Digital Resources: An On-the-Ground View of Projects Today. http://www.sr.ithaka.org/sites/default/files/reports/4.17.1.pdf (accessed September 2015).

Maron N. L. and Picke S. (2014). Sustainability Implementation Toolkit. Developing an Institutional Strategy for Supporting Digital HumanitiesResources". http://www.sr.ithaka.org/publications/sustainability-implementation-toolkit/ (accessed October 2015).

Ullyot, M. (2013). Digital Humanities Projects, *Renaissance Quarterly*, **66**(3): 937–47. doi:10.1086/673587.

Villa, C. (2015). Project Sustainability in DH: Collaboration and Community. http://digitalhumanities.berkeley.edu/blog/15/05/01/project-sustainability-dh-collaboration-and-community. (accessed October 2015).

Van Zundert, J. (2012). If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities, *Historical Social Research Journal*, **37**(3).

Warwick, C., Terras, M., Huntington, P. and Pappa, N. (2006). *If you build it will they come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data*. Selected papers from Digital Humanities 2006, Paris-Sorbonne.

Webb, S. (2015). Requirements and National Digital Infratsructures: Digital Preservation in the Humanities. http://breac.nd.edu/articles/61504-requirements-and-national-digital-infrastructures-digital-preservation-in-the-humanities/ (accessed October 2015).

Zhang, Y. (2010). Developing a Holistic Model for Digital Library Evaluation, *Journal of the American Society for Information Science and Technology*, **61**(1): 88–110. doi:10.1002/asi.21220.

Zorinch, D. Z. (2003). *A Survey of Digital Cultural Heritage Initiative and Their Sustainability Concerns*, Council on Library and Information Resources Washington, D. C. http://www.clir.org/pubs/reports/pub118/pub118.pdf (accessed August 2015).

Notes

[1] http://www.apex-project.eu/index.php/en/
[2] https://textgrid.de/
[3] http://www.cendari.eu/
[4] https://www.dariah.eu

If You Build It Will They Come? Digital Infrastructure And Disciplinary Practice In Language Documentation

Simon Musgrave
simon.musgrave@monash.edu
Monash University, Australia

Nick Thieberger
thien@unimelb.edu.au
University of Melbourne

Digital scholarship depends on the availability of data in forms which are tractable to computational techniques, implying storage of data in sustainable archives. This is perhaps even more true of research in the humanities than in science. As the ESF observed in their 2011 report on research infrastructure: "in the hard sciences, datasets tend to be generated rather than collected and tend to be homogeneous in nature .... In Humanities, data tends to be collected and to be heterogeneous in content and format" (European Science Foundation, 2011:5). It is both the possibility of storing the data and the stored data itself which makes up critical infrastructure in many disciplines and innovative scholarly practices may not develop in the absence of such infrastructure. Here, we discuss the development of digital infrastructure in the field of language documentation (within the discipline of linguistics) and try to assess the extent to which the provision of well-funded (by the standards of the discipline) infrastructure is changing scholarly practice.

Language documentation as a field in linguistics dates from the publication of the seminal paper 'Documentary and descriptive linguistics' (Himmelmann, 1998). It is in large part a reaction by linguists to the challenge of language endangerment (see Hale et al., 1992; and Musgrave, 2015 for a recent survey) and it emphasises the collection of large bodies of data of languages in use. Language documentation began at the same time that it was becoming feasible to make high quality digital recordings, both audio and video, on reasonably priced equipment. Archiving of documentary material was a core component of the program as conceived by Himmelmann, and this was immediately seen to mean digital archiving. Indeed, it can be argued that the whole enterprise of documentary linguistics falls comfortably within the digital humanities (Thieberger, 2014).

Two projects began in the first decade of the century to fund researchers to make collections of documentary materials. One was based at the Max Planck Institute for Psycholinguistics (Nijmegen, The Netherlands) and funded by the Volkswagen Stiftung;[1] the other was based at the School of Oriental and African Studies (London, UK), funded by Arcadia.[2] MPI Nijmegen had a well-resourced technical department which took responsibility for developing the archiving stream of the DoBeS project. HRELP had to build their archive from scratch; no existing resources at SOAS were available to support their work. For both projects, it was a requirement of funded research that data were deposited in the relevant archive. Both archives, however, are open to deposits from non-funded researchers.

The DoBeS archive is subsumed under a larger archive called The Language Archive[3] which also holds language data from other MPI activities and from other institutions and projects. Figure 1 shows the cumulative deposits in TLA for DoBeS material and for material classified as donated. The figures here represent the number of files retrieved from the catalogue based on their 'Last modified' field. This does not reflect the accession date of the file in all cases, but it is a reasonable proxy.[4]

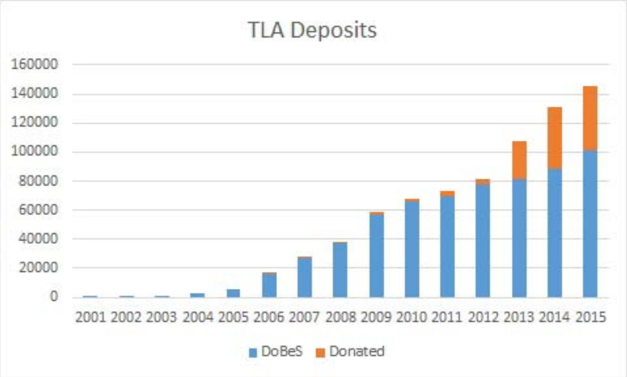


Figure 1. Cumulative deposits in TLA

The archive at SOAS is known as ELAR. Figure 2 shows cumulative deposits in that archive; figures are drawn from information provided by ELAR in the annual reports of HRELP and represent bundles of data deposited rather than individual data files.

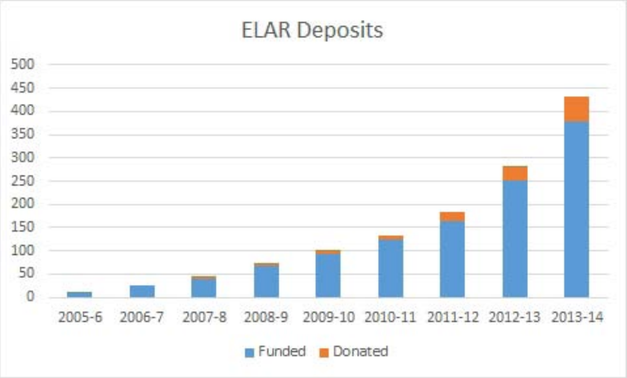


Figure 2. Cumulative deposits in ELAR

Although both archives have grown over the periods shown, the patterns are different, with a flattening out in the DoBeS deposits starting around 2011. There was no funding round in 2010. Projects were funded in 2011 and 2012 but 2012 was the final round and while the archive still expects to receive deposits from the last two funding rounds, such deposits will continue to decrease.

Figure 3 shows the percentage of donated deposits based on the cumulative figures for each archive.

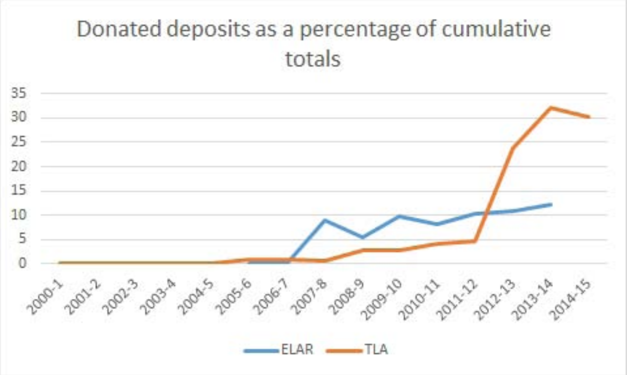


Figure 3. Comparison of the percentage of donated material in the two archives

For both archives, once the infrastructure was established, non-funded deposits make up an increasing proportion of the archive. TLA has added large amounts of donated material in recent years, in part a conscious effort by the archive to expand its activities beyond the DoBeS project by taking responsibility for existing data sets. One barrier to archiving data is meeting an archive's requirements particularly in the area of metadata. ELAR uses a more flexible metadata system than TLA (which uses the IMDI scheme, https://tla.mpi.nl/imdi-metadata/). We might therefore expect more voluntary deposits in ELAR than in TLA, but this is not the case in these data. The different proxies we are using to assess these trends here make it difficult to compare the two exactly, but we can see a clear trend over the last decade which suggests that archiving data is increasingly a part of scholarly practice in this area of linguistics and that there has been progress since the rather gloomy summary provided by Thieberger (2011).

One question raised by this data is the extent to which the donated data is coming from researchers who have also been funded by the relevant program. Both projects provide training to funded researchers and it is possible that the practices learned there are continued when researchers collect data in other projects. In the case of TLA, 36 data sets have been deposited which were not the result of DoBeS funding and in six cases the researcher(s) had been funded by DoBeS for other work. ELAR has material not associated with funded projects deposited by 37 researchers of whom eight had been funded by ELDP for other work. These figures suggest that acquiring experience of the archiving process is a factor in future work practices, and that this factor has had very similar levels of effect in both archives.

The data which we have used in this paper are limited and important questions remain unanswered. The most obvious is how and to what extent are these resources being used. ELAR provide some usage statistics; based on logs for the month of November 2013, estimated traffic on the catalogue and the archive portal is around 680 users per day with an estimated 1.66 million pages served per year. These numbers suggest that the archive is being used a lot, but it is not possible to tell who the users are or what they are doing on the site. More fine-grained data are needed to tell us whether the availability of resources such as ELAR and TLA is changing scholarly practice in accessing language data.

The data we present here indicates that, in the field of documentary linguistics, the availability of good infrastructure for digital archiving has had an impact on scholarship. Data is being deposited in the archives beyond the requirements of the relevant funding bodies, and in the case of the archive for which usage statistics are available, it seems that the level of activity is substantial. These changes may in part be due to the level of training provided to funded researchers; but such training should, we suggest, be considered an essential part of the research infrastructure of digital scholarship. Another reason for the changes may be that linguistic scholarship is moving towards recognising primary data as scholarly output (Thieberger et al., 2016). This process is only possible when primary data can be cited using persistent identifiers provided by a repository (cf. NSF (Task Force on Data Policies), 2011:9 (Recommendation 2)). There is therefore the possibility of a virtuous circle here: adoption of best practice in data management will lead to more robust research methods in the field as well as career benefits for researchers.

Bibliography

European Science Foundation (2011). *Research Infrastructures in the Digital Humanities*. (Science Policy Briefing). Strasbourg: European Sceince Foundation http://www.esf.org/fileadmin/

Public_documents/Publications/spb42_RI_DigitalHumanities.pdf (accessed 2 March 2015).

Hale, K., Krauss, M., Watahomigie, L. J., Yamamoto, A. Y., Craig, C., Jeanne, L. M. and England, N. C. (1992). Endangered languages. *Language*, pp. 1–42.

Himmelmann, N. (1998). Documentary and descriptive linguistics. *Linguistics*, **36**(1): 161–96.

Musgrave, S. (2015). Endangered languages. In Allan, K. (ed), *The Routledge Handbook of Linguistics*. (Routledge Handbooks in Linguistics). Milton Park, Abingdon, Oxon ; New York, NY: Routledge, pp. 385–400.

NSF (Task Force on Data Policies) (2011). *Digital Research Data Sharing and Management*. Washington DC: National Science Foundation.

Thieberger, N. (2011). Where are the records? *Endangered Languages and Cultures*. http://www.paradisec.org.au/blog/2011/06/5649/ (accessed 1 November 2015).

Thieberger, N. (2014). Digital humanities and language documentation. In Gawne, L. and Vaughan, J. (eds), *Selected Papers from the 44th Conference of the Australian Linguistic Society, 2013*. Melbourne: University of Melbourne, pp. 144–59. http://minerva-access.unimelb.edu.au/handle/11343/40961 (accessed 30 October 2015).

Thieberger, N., Margetts, A., Morey, S. and Musgrave, S. (2016). Assessing Annotated Corpora as Research Output. *Australian Journal of Linguistics*, **36**(1): 1–21. doi:10.1080/07268602.2016.1109428.

Notes

[1] Dokumentation Bedrohte Sprache (DoBeS): http://dobes.mpi.nl/

[2] Hans Rausing Endangered Languages Project: http://www.hrelp.org

[3] https://tla.mpi.nl

[4] We are deeply grateful to Paul Trilsbeek of The Language Archive for assistance in refining this data.

The Dialogic Turn and the Performance of Gender: the English Canon 1782-2011

Grace Muzny
muzny@stanford.edu
Stanford Universtiy, United States of America

Mark Algee-Hewitt
malgeehe@stanford.edu
Stanford Universtiy, United States of America

Dan Jurafsky
jurafsky@stanford.edu
Stanford Universtiy, United States of America

Understanding how the spoken language is represented in novels over time, and how this relates to gender and other characteristics of the represented speaker and the author is a key question in the Digital Humanities. Previous work has explored such questions lexically, focusing for example on differences in word choice between male and female authors. Yet while such lexical stylistic approaches have been computationally sound (Argamon 2003, Olsen 2005, Yu 2014, Rybicki 2015), a purely lexical approach is known to have serious methodological dangers. Individual lexical items are highly conflated with topic, genre, author idiosyncrasies, and era, making it difficult to draw general conclusions, particularly over long time periods. An even greater problem is that this approach essentializes gender a priori (Bing & Bergvall 1996), neglecting how the complex interplay between an author and the characters they portray creates the performance of linguistically gendered writing.

We propose to draw from methods in social psychology (Newman 2008), network analysis (Schwartz 2013), and corpus linguistics (Biber 1991) to offer two innovations in the analysis of novels. The first is a new metric for characterizing dialogue, called *dialogism*, that references Bakhtin's theories of novels and dialogue (1935). This measurement uses abstract grammatical features in the text to characterize the extent to which it is dialogic. By using categories like parts of speech, our method avoids the genre-specific and era-specific problems of individual words. Our second innovation is a computational analysis of the performance of gender within dialogue. We explore the relationship between the gender of the characters portrayed in novels and their language, which illuminates the performative aspects of gender.

Although we do not entirely escape the entrapments of gender as a defining authorship trait in our analysis, we do use it to develop deeper understanding of dialogue as a whole rather than simply a tool of reifying stereotypes. From our innovations, we propose to answer three important questions: 1) what is the composition of the dialogic landscape in fiction? 2) what characterizes dialogue linguistically? and 3) how is author gender and the gender performed by characters reflected in the language of dialogue?

Data and Dialogue Extraction

Using three corpora of novels from the beginning of the romantic period to the present day, we construct a corpus that contains 1,106 novels from 1782 until 2011. This corpus is largely composed of the standard English canon, with the pre-1900 portion drawn from Chadwyck-Healey, and therefore exhibits the bias present in canonical authorship, containing 851 male-authored novels and 255 female-authored novels.

We introduce and use a new dialogue extraction system to locate quoted text based on a series of rules and regular expressions.Then, using a number of high-precision patterns such as <QUOTE>-<PRONOUN>-<VERB>, we assign speaker gender to quotes associated with gender-disambiguating pronouns (*he, she,* etc.) or proper names that can be reliably distinguished from gendered namelists. We evaluate our system on texts hand-labeled for quotes and speaker identity (Austen's *Pride and Prejudice* (He 2013), Cooper's *The Spy*, and Fitzgerald's *The Great Gatsby*), resulting in average quote extraction of 94.5% of quotes at 95.4% precision and gender attribution of an average of 44.4% of the extracted quotes at 93.1% precision.

The Dialogic Landscape

We first statistically evaluate the distributions of extracted dialogue to see whether novels have become more dialogue driven over time. Controlling for novel length (Figure 1), we measure mean quotes per thousand words per decade, finding a steady increase over time regardless of author gender (Figure 2). Overall, male authors add roughly 1 quote per 20 years (r^2 =.80); female authors 1 quote every 30 years (r^2 = .64).

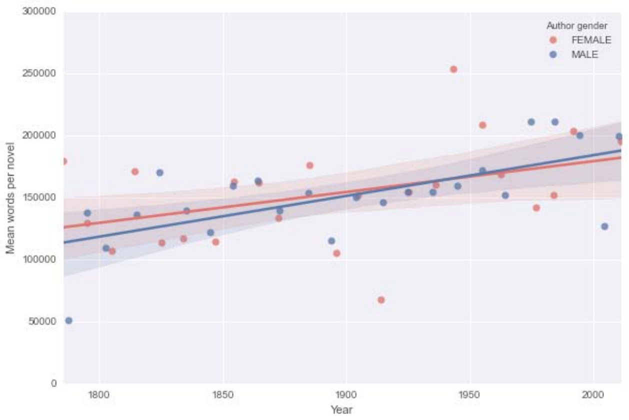


Figure 1: Average novel length over time

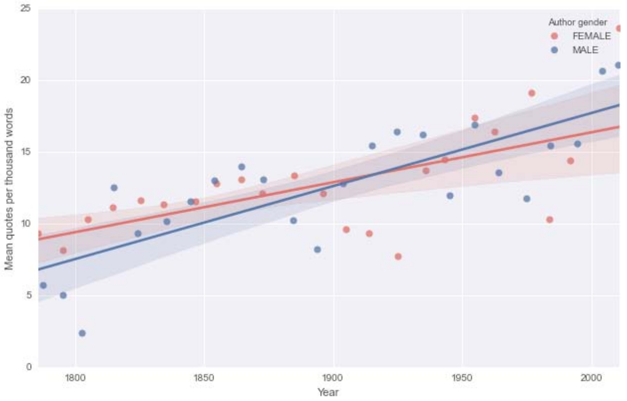


Figure 2: Quotes per thousand words by author gender over time

Next, we examine the relative attention that authors pay to characters and gender, using dialogue as our lens. From the extracted speaker-assigned quotes we calculate the proportion of male- to female-spoken dialogue per novel. This allows us to understand the gender of the characters portrayed in our texts and how this composition has changed over time.

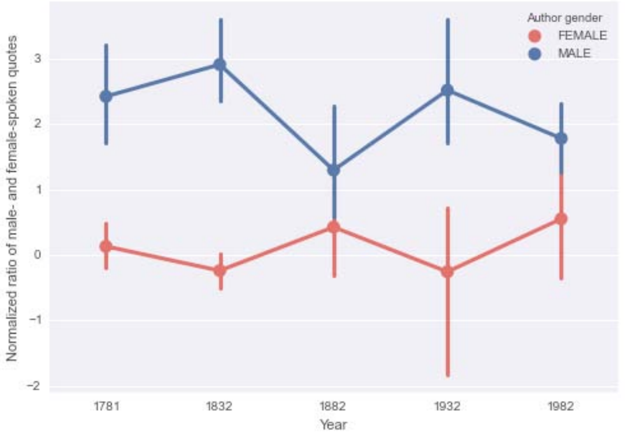


Figure 3: Mean normalized ratio of male- and female-spoken quotes by novel over time by author gender, bucketed by 50 year intervals

We compute the mean of a normalized ratio of words spoken by male versus female characters per novel, bucketed by 50 year intervals (Figure 3). This ratio is normalized such that it forms a continuous spectrum centered at 0, with +1 signifying that male characters spoke twice as often; -1 that female characters did.

Here, we see that male authors tend to write male-spoken dialogue and female authors female-spoken dialogue, but that overall, female-authored dialogue tends to be close to balanced (mean = -.04) while male-authored dialogue is far from it (mean = 2.5). An interesting leap towards balanced portrayal happens among male authors at the beginning of the 20th century, influenced largely by authors such as Henry James, E.M. Forster, and Booth Tarkington.

Linguistic Characterization of Dialogue

To discover the underlying linguistic differences between narration and dialogue, we perform Multi-Dimensional Analysis (MDA) on the dialogue in our corpus, using a slightly modified set of Penn Treebank part-of-speech tags as distinguishing features. While it is not a classification algorithm, MDA isolates factors such that data with similar features are grouped together. These factors contain features that are positively or negatively correlated with one another. Effectively, even though no class labels are used, these factors reveal strong feature relationships for dialogic text.

Based on this analysis we propose a new dialogue metric, *dialogism*, that is robust to lexical choice and transportable across corpora. The new metric is constructed from the ten factors (out of 16) with $r^2 > 0.5$ when used to separate narration from dialogue and considers both positive indicators and negative indicators of dialogue, shown in Table 1. Figure 4 shows the high overall separation between narration and dialogue that our metric achieves (Kolmogorov-Smirnov value = 0.89).

+	-
Present tense verbs, bare verbs, modals, 1st/2nd/"it" pronouns, Wh-pronouns, interjections, existential there, adverbs	past tense verbs, 3rd person pronouns, gerunds, particles, nominalizations, determiners, Wh-determiners, prepositions and subordinating conjunctions, adjectives

Table 1: Positive and negative contributors to dialogism score

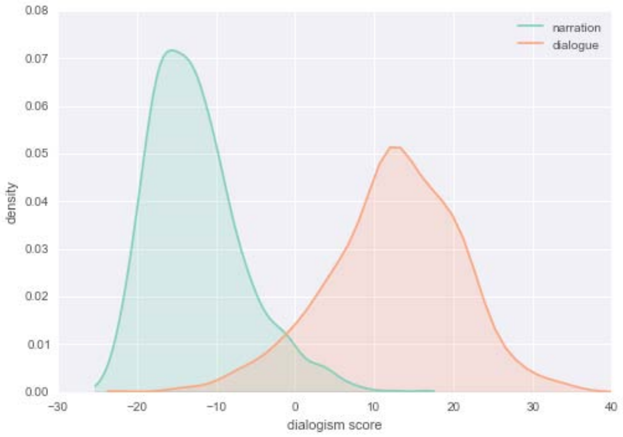


Figure 4: dialogism scores for narration and dialogue

Examing dialogism over time, we find that the distance between narration and dialogue has increased since the 18th century, and that, as a whole, novels are becoming more dialogic. At the same time, visualizing the outlying data quantifies the effects that shifts in literary style, for instance, modernism in the early 20th century, had in terms of dialogism.

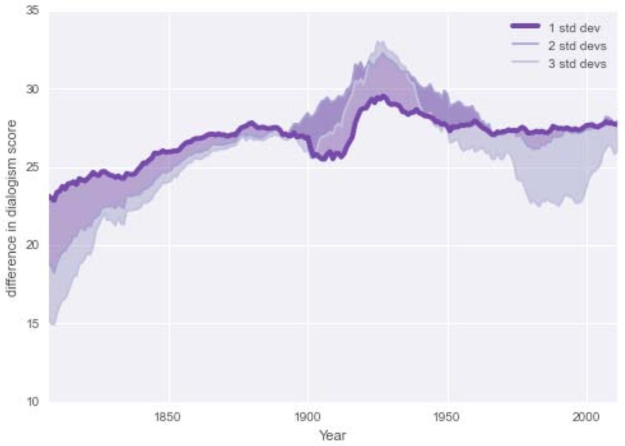


Figure 5: Difference in dialogism scores for dialogue and narration over time for texts within 1 (73.1%), 2 (93.5%), and 3 (99.5%) standard deviations of the mean difference.

Gender in Dialogue: Performance and Authorship

We finally turn to the performative aspects of gender: how do authors perform gender through the speech of their gendered characters? Which authors most significantly differentiate their male and female characters through dialogue?

Using the same corpus, but subsampling to balance for speaker gender, we isolate author gender effects and use these results combined with the original data to isolate speaker gender effects. This analysis reveals that the differences between male- and female-authored dialogue cannot be accounted for solely based on either author gender or on the genders of the speakers they portray. Author gender accounts for 64% of the variation between male- and female-authored dialogue while other speaker gender effects account for 36%. However, because the standard deviation of this ratio is so high (23%), this is an indication that some authors are more heavily influenced by their own genders and some more by the characters they portray.

Digging deeper into the question of which authors are better than others at differentiating male and female characters through dialogue, we perform t-tests on the male- versus female-spoken quotes for each author. Shown in Figure 6 are the authors who differentiate their male and female characters through dialogue at $p < .0005$. Notably, while most authors portray female characters as more dialogic in speech, a small minority do the opposite (above, Maria Cummins), a trend that also holds at higher p thresholds.

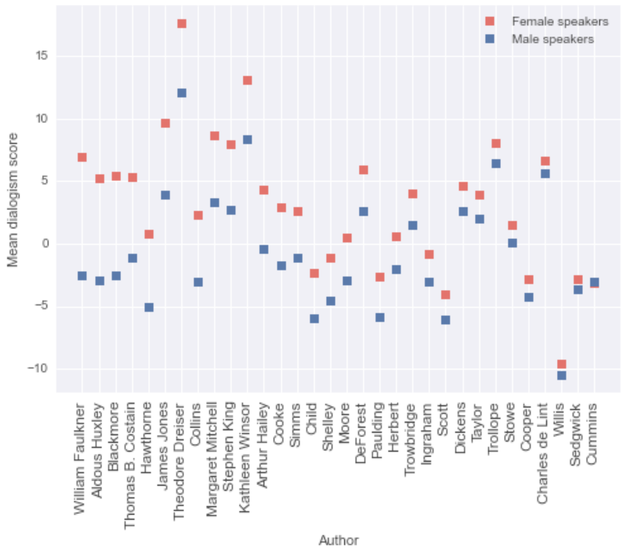


Figure 6: Authors whose male and female characters are significantly differentiated by dialogism score at p < .0005

Conclusion

The development of the novel is marked by an increased use of dialogue over time. This suggests that the deepening of characters was accompanied by, or may be an effect of, a shift of attention towards performative modes of characterization. Moreover, this transformation to a more dialoguedriven structure bears a gendered dimension, suggesting that the depiction of sociality by female novelists favored a more realistic gender balance than the predominately male social models favored in maleauthored texts. Our work suggests that the presence of any gendered language in the text may be contingent upon the mode of performance adopted by the author, regardless of gender, an observation supported by the variation in both dialogue cast composition and dialogism at the beginning of the 20th century, when a shift in literary style occurred. Further, the strict delineation of midnineteenth century sociality into the gendered public and private spheres, as represented by the novel, is itself a deeply gendered understanding of sociocultural codes more true of maleauthored texts than femaleauthored ones. The substantial effects that speaker gender has on dialogue indicates that perhaps not only are novels intrinsically dialogic, but that dialogue itself is intrinsically performative. Thus, the performative nature of a novel is itself deepened by the degree to which its' author differentiates the characters through a gendered performance of dialogue.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-114747. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank their anonymous reviewers, the Stanford LitLab, and the Stanford NLP group for their helpful feedback.

Bibliography

Argamon, S., et al. (2003). Gender, genre, and writing style in formal written texts. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN*, **23**(3): 321-46.

Bakhtin, M. M. (1935). Discourse in the novel. *The Novel: An Anthology of Criticism And Theory 1900–2000*, pp. 481-510.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Bing, J. M. and Bergvall, V. L. (1996). The question of questions: Beyond binary thinking. *Rethinking language and gender research: Theory and practice*, **1**: 30.

He, H., Barbosa, D. and Kondrak, G.. (2013). Identification of Speakers in Novels. *ACL*, (1).

Newman, M. L. et al. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, **45**(3): 211-36.

Olsen, M. (2005). Écriture féminine: searching for an Indefinable practice?. *Literary and linguistic computing*.

Rybicki, J. (2015). Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies. *Digital Scholarship in the Humanities*. fqv023.

Schwartz, H. A., et al. (2013). "Personality, gender, and age in the language of social media: The open-vocabulary approach." *PloS One* **8**(9): e73791.

Yu, B. (2014). Language and gender in Congressional speech. *Literary and Linguistic Computing*, **29**(1): 118-32.

Digital Humanities in Cultural Areas Using Texts That Lack Word Spacing

Kiyonori Nagasaki
nagasaki@dhii.jp
International Institute for Digital Humanities, Japan

Toru Tomabechi
tomabechi@dhii.jp
International Institute for Digital Humanities, Japan

Charles Muller
acmuller@l.u-tokyo.ac.jp
The University of Tokyo

Masahiro Shimoda
shimoda@l.u-tokyo.ac.jp
The University of Tokyo

As most of modern and pre-modern western writing systems explicitly represent division of the words in a sentence by spaces or breaks, it has been easy to use computers to analyze texts based on each word and its meanings. However, there are several modern and pre-modern writing systems that do not explicitly indicate word separation in texts; that is, all words in a sentence are contiguous. A major contemporary representative of this kind of writing is seen in the language system of East Asia. Moreover, a popular Japanese pre-modern writing system called kuzushi-ji (cursive style characters) had often been presented with undivided characters even in typesetting until the late nineteenth century (Fig. 1, Fig. 2). As the lack of word-separation has been evoking not only ambiguity but also multiple interpretations, it has formed an aspect of cultural richness in Japanese culture. However, as a result, Japanese texts have intrinsically presented difficulties: not only in the case of textual analysis but also in both manual and automatic transcription in the digital era. This presentation will discuss problems in these writing systems and the current situation of attempts to resolve them through the methods of digital humanities.

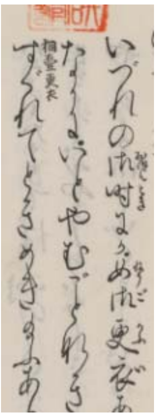


Fig.1: Typesetting printing (*Ise-monogatari* (Tales in Ise). Saga-bon. 1608. http://dl.ndl.go.jp/info:ndljp/pid/ 1287963/6)

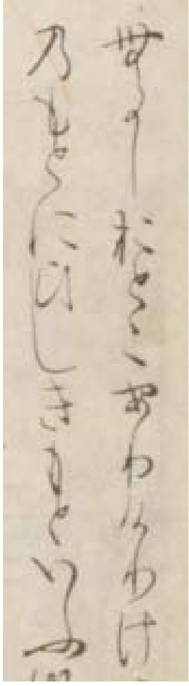
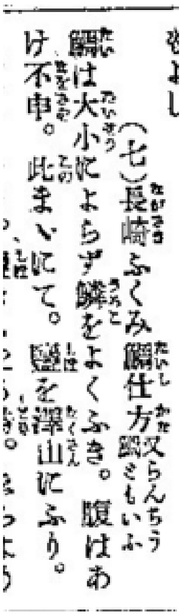


Fig. 2: Woodcut printing (Yamamoto Shunsho ed. *Eiri-Genji-Monogatari* (Pictorial tales of Genji. 1654. http://base1.nijl.ac.jp/~anthologyfulltext/)

Difficulties of transcription

Recent Japanese texts do not have serious problem in case of OCR due not only to the separation of each character but also accuracy and clarify of its printing. However, it is difficult to OCR books printed even ten decades ago because of two points: most of them uses relatively complicated characters for OCR and parallel embedded small-font size texts (called ruby in HTML5) which explain pronunciation of a word, and are too close to the explained word to OCR (Fig. 3), even though they were printed by metal typesetting. More three decades ago characters were sometimes connected, and the writing style of characters were partially cursive (Fig. 4). Recently, some researchers are attempting to develop tools for recognition of kuzushi-ji not based on the shape of individual characters but by continuous shapes of characters. They have not yet reached the stage where they are able to transcribe all characters accurately, for both technical and intrinsic reasons, but the technology can nonetheless assist in reading such texts by showing candidates of characters (Fig. 5)[1]. One of reasons why such kind of image recognition of a series of characters by machine is available in many cases is that many resources written by kuzushi-ji are woodcut printing, in which case the continuous cursive characters are more or less normalized within a single book.



をよし
（七）長崎ふくみ鯛仕方又らんちう鯛ともいふ
鯛は大小によらず鱗をよくふき。腹はお
け不申。此まゝにて。鹽を澤山にふり。

Fig. 3: Ruby close to text body (Ohashi Matatarou. *Jituyou-ryouri-hou,* A Guidebook to practical cooking). Hakubunkan. 1895. http://dl.ndl.go.jp/info:ndljp/pid/849051/19

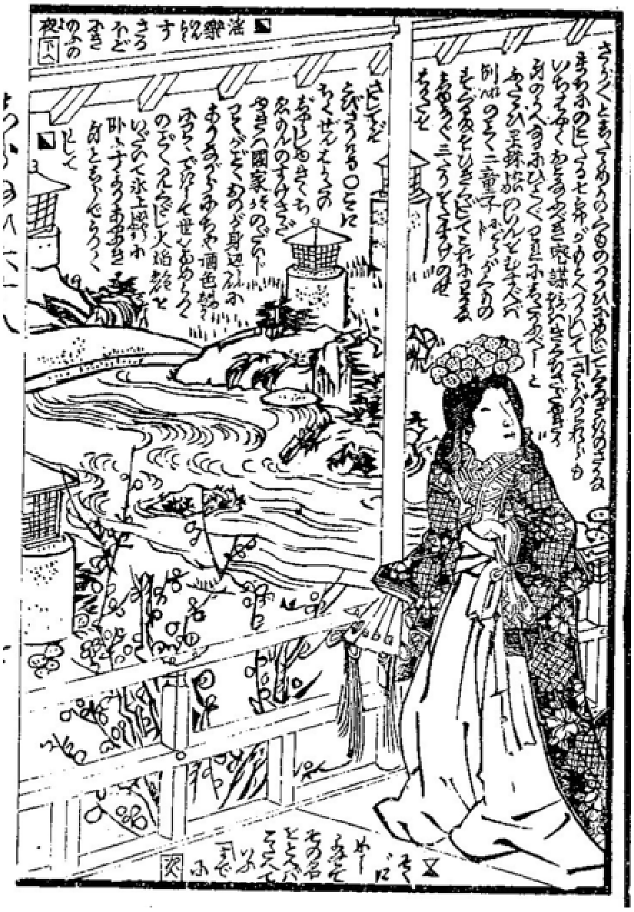


Fig. 4: Continuous characters (Ryusuitei Tanekiyo ed. *Shiranui-monogatari* (Tales of Shiranui). Vol. 68a. Enju-dou. 1885. http://dl.ndl.go.jp/info:ndljp/pid/ 884924/8)

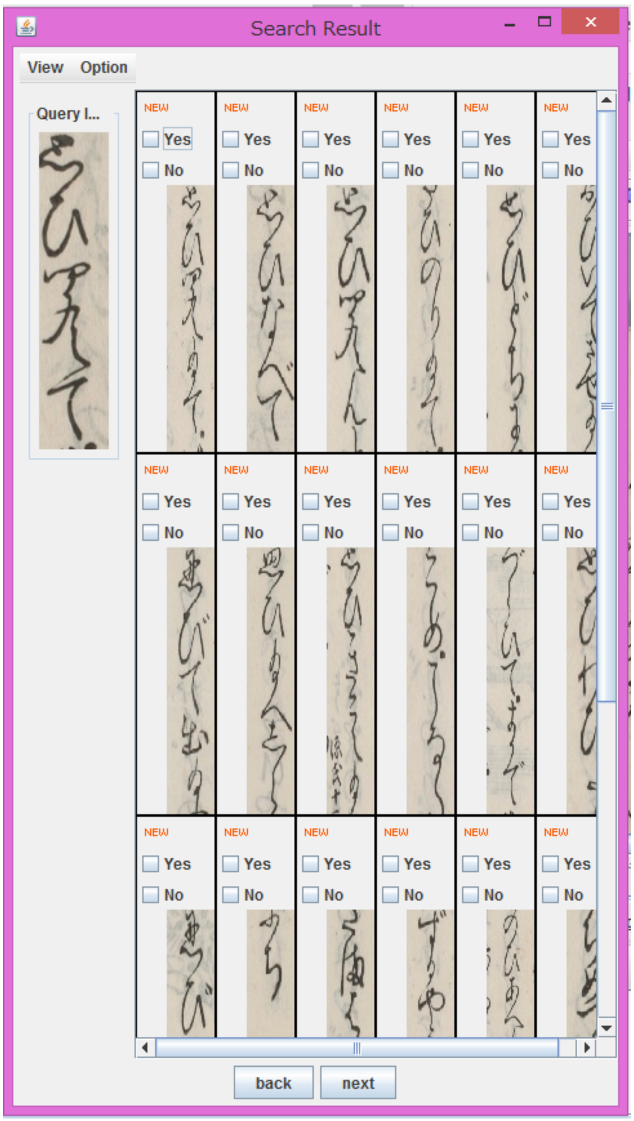


Fig. 5: A result of image search in SMART-GS

However, there are special difficulties presented when a needed character is not encoded in Unicode. It seems to be similar with the case of Medieval Unicode Font Initiative[2], but the number of unencoded characters would be much more in the Japanese case included in East Asia culture. Especially, as Japanese culture has been involved with foreign cultures and developing them in its contexts, several writing systems are preserved in its cultural resources, including Kanji, Hiragana, Katakana, Hentaigana, and Siddham scripts. Siddham scripts were encoded in Unicode 8.0 with its variant characters by efforts of Script Encoding Initiative, international experts, and SAT project[3]. There are already 80,000 Kanji (CJK unified characters) registered, but thus number will continue to increase. Hentaigana (including over 200 glyph shapes) was proposed to the ISO committee on October 2015[4]. In order to make easy-use digital scholarly edition for Japanese texts, especially classics, this process will be continued.

While efforts of transcription, due to commoditization of digitizing textual materials in hi-resolution, digital image databases have also been grown in Japan. Especially, the National Diet Library in Japan has been addressing the publication of digitized collection including over 300,000 books--since over decades ago and recently stated that most of them are to be released in the public domain[5]. And some institutes such as Kyoto Prefectural Library and Archives[6] and the University of Tokyo Library[7] are publishing their digitized collections under open license. The Art Research Center in Ritsumeikan University and the National Institute of Japanese Literature[8] have released many digitized textual resources under academic license. The latter institute plans to distribute parts of their contents under open license in this year in its new comprehensive digitization project[9]. Needless to say, these are useful to enhance the convenience of humanities research. Especially, in Japanese contexts, many humanities researchers mention that validation of research results has been made much more efficient by the increased use of the digitized images.

Crowd sourcing transcription has recently emerged also in Japan. Transcribe JP project has been conducted as a SIG of the Japanese Association for Digital Humanities. It provides a Web service[10] for transcription with Omeka and Scripto plugin. Moreover, it started a micro task crowd sourcing project on[11] October 2015 in cooperation with Crowd4U project[12]. Contributors can determine whether a character is exactly OCRed or not, comparing a candidate character with a piece of an image only by one click. The first experiment was finished in a much shorter time than we expected. Further results will be reported at the DH2016.

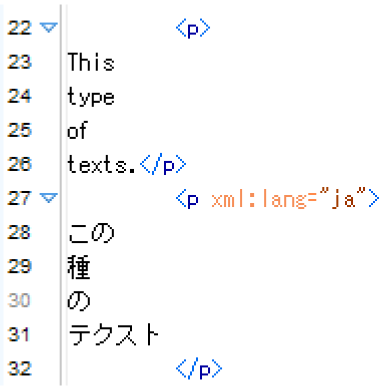
Difficulties in Word Separation

In spite of the difficulties of transcription, there are many digitized texts in Japanese. Aozora-Bunko[13], a public domain Japanese texts repository similar to the Gutenberg Project, provides over 10,000 texts on its Web site and GitHub. The National Institute of Japanese Language and Linguistics (NINJAL)[14] publishes several encoded historical Japanese texts with POS tags on Web and Web services of textual analysis on modern Japanese texts including 100 million words with POS tags each word in its original format. The SAT project[15] also provides digital texts of Buddhist scriptures consists of 100 million characters mainly in Chinese and Japanese with some philological tags on Web.

The texts of NINJAL consists of separated words with POS tags, but most of the others do not use this method. Then, methods for textual analysis are common in Japan: The one is n-gram analysis regarded a character as one "n". The other is developing tools for automatic separation of words sometimes with POS tagger, such as Mecab[16], Chasen[17], and Kuromoji[18]. These tools realize a high degree of precision, but sometimes produce erros. In this case, one has to manually correct the result of the tools if sharing exactly-processed texts is necessary. Moreover, even if a separation is not mistaken, it might support an interpretation in some cases. Such kinds of cases can also be occurred in word-separated corpora. This type of writing system includes such kinds of issues.

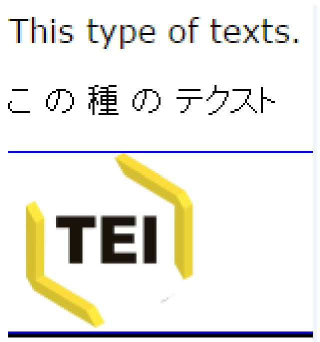
Rendering of texts

In XML-formatted texts, suc has those maintained in TEI, JATS[19], and so on representation of breaks in source XML files seems to be regarded a space as a separation between words in popular stylesheets. But in the case of non-separated texts, it causes problems such as unnecessary separation. The XSLT-processed Japanese text in fig.7 must exclude spaces between characters in spite of line-breaks in the XML source (fig.6). Conversely, as a Japanese semi-governmental open access journal system adopting JATS ignores line breaks even in English, the words are connected in the case of Fig. 6 and Fig.7. This problem seems to be recognized in ePub with solution in CSS according to the target language[20]. While it must already be discussed even in contexts of DH because non-spacing texts have been generated in various time and place, the differences of treatment of the line-breaks in XML source files should be carefully treated regarding not only representation but also analysis of texts.



```
            <p>
This
type
of
texts.</p>
            <p xml:lang="ja">
この
種
の
テクスト
            </p>
```

Fig. 6 An example for contrast of word separation in XML format



This type of texts.

こ の 種 の テクスト

Fig. 7 An example for a result of XSL Transformation of the Fig. 6

In contexts of current DH, huge humanities resources have still been dormant. According to their awakening, these kind of issues should be gradually revealed and needed to be solved from both practical and abstract viewpoints. Through solving them earnestly under global communication, DH will come to better fruition.

Notes

[1] Hashimoto, Yuta, et al. The SMART-GS Project: An Approach to Image-based Digital Humanities. *Digital Humanities 2014*:476-477. 2014.
[2] http://folk.uib.no/hnooh/mufi/
[3] Pandey, Anshuman. Proposal to Encode the Siddham Script in ISO/IEC 10646. ISO/IEC JTC1/SC2/WG2 N4294. 2012. http://www.unicode.org/L2/L2012/12234r-n4294-siddham.pdf . KAWABATA , Taichi, Toshiya SUZUKI, Kiyonori NAGASAKI and Masahiro SHIMODA. Proposal to Encode Variants for Siddham Script. ISO/IEC JTC1/SC2/WG2 N4407. 2013. http://std.dkuug.dk/JTC1/SC2/WG2/docs/n4407.pdf .Anderson, Deborah, et al. 2013-11-22 Siddham Script (梵字) Meeting @ Tokyo, JAPAN, Earth. ISO/IEC JTC1/SC2/WG2 N4523. 2013. http://std.dkuug.dk/JTC1/SC2/WG2/docs/n4523.pdf .
[4] ITSCJ SC2 Committee, IPSJ, JAPAN. Proposal of Japanese HENTAIGANA. ISO/IEC JTC1/SC2/WG2 N4674. 2015. http://unicode.org/wg2/docs/n4674-Japan_Hentaigana_Proposal-a.zip .
[5] http://dl.ndl.go.jp/
[6] http://hyakugo.kyoto.jp/
[7] http://dzkimgs.l.u-tokyo.ac.jp/utlib_kakouzou.php
[8] http://www.nijl.ac.jp/
[9] http://www.nijl.ac.jp/pages/cijproject/index_e.html
[10] Hondigi2014. http://lab.ndl.go.jp/dhii/omk2/
[11] 翻デジ@JADH×Crowd4U. http://www.jadh.org/transcribejp
[12] Crowd4U. http://crowd4u.org/en/
[13] http://www.aozora.gr.jp/
[14] http://www.ninjal.ac.jp/
[15] http://21dzk.l.u-tokyo.ac.jp/SAT/
[16] http://taku910.github.io/mecab/
[17] http://chasen.naist.jp/hiki/ChaSen/
[18] http://www.atilika.com/ja/products/kuromoji.html
[19] http://jats.nlm.nih.gov/
[20] http://www.idpf.org/epub/30/spec/epub30-overview.html

Player-Driven Content: Analysing Textual Communications in Online Roleplay

James O'Sullivan
josullivan.c@gmail.com
Pennsylvania State University

Michelle Shade
mas746@psu.edu
Pennsylvania State University

Ben Rowles
blr5241@psu.edu
Pennsylvania State University

Introduction

The purpose of this study is to determine the extent to which online roleplayers[1] make use of language in the construction of narrative. Using computational approaches to text analysis, we compare in-game chatlogs of roleplayers with those of non-roleplayers. In doing so, we identify the particularities of the language of roleplay. Our findings are significant in that they macro-analytically demonstrate the differences between the narrative language of roleplayers and the objective-driven language of traditional players. While these differences have already been analyzed by other researchers, this study is the first to offer a thorough account of them using quantitative methods.

Roleplay is both a historic and present practice, dating as far back as ancient Greece (Corsini, Shaw and Blake, 1961). Throughout the 1970s and 80s, face-to-face roleplay surged in popularity with the advent of renaissance fairs and tabletop games such as *Dungeons and Dragons* (Barton, 2008). Today, MMORPGs (Massive Multiplayer Online Roleplaying Games)[2] are believed to enjoy over 47 million collective subscriptions. Research on the communication habits of online roleplayers provides insight into the contemporary refiguring of this long-running narrative and representational practice. The existing research on roleplay in video games relies on qualitative methods and focuses on games that have since waned in popularity. Our game of choice is *World of Warcraft*, commercially one of the world's most popular MMORPGs (Bainbridge, 2015), and thus a rich and legitimate source of data. We approach the analysis of this dataset using a range of computer-assisted methods.

Methodology and Results

For the purposes of this study, we gathered two sets of data:

- Chatlogs volunteered by *World of Warcraft* roleplayers

- Guild A donated strictly-roleplay chatlogs and non-roleplay chatlogsGuild B donated strictly-roleplay chatlogs
- Chatlogs volunteered by *World of Warcraft* non-roleplayers
- Guild C donated non-roleplay chatlogs from a server without background roleplayNon-RP RP server consisted of the study authors' chatlogs, from a server with background roleplay mixed with non-roleplay (task-oriented gameplay and casual chat)
- this sample provided a middle ground between strictly-roleplay and strictly non-roleplay samples

We applied a variety of Digital Humanities methodologies to our dataset in an effort to extrapolate the differences in language of interest to this study. Our experiments included:

- Most frequent (uncommon) word analyses to determine if any dominant themes differ across roleplayers and non-roleplayers.

As can be seen (see Fig. 1), the majority of words in the non-roleplay sets are related to objective-based gameplay, whereas in the roleplay sets, there is a dominance of emotive and descriptive words relating to interactions between characters.

- Delta analysis to determine if chatlogs clustered by style depending on whether they were taken from roleplay or standard play.

We were able to obtain both roleplay and non-roleplay chatlogs from the same group of players (Guild A), allowing us to perform a valid stylometric analysis (see Fig. 2) to determine if there is a stylistic separation between the language used in roleplay and non-roleplay. We conduct the Delta analysis using R, which shows that the linguistic styles of these player groups are distinct.

- Zeta analysis to establish words that were distinctive to each player group.

Our Zeta analysis reveals those words which are distinct to each group. Words particular to the non-roleplayers include objective-based terms such as "dps" (damage per second), and slang terms such as "nerf" (overkill). Meanwhile, the roleplayers tend to use descriptors of body language (e.g. "looks", "nods","stares", "glares" and "grins") and otherwise emotive terms (e.g. "cheers", "breath", "growls").

- Topic modelling to determine any discursive trends across roleplayers and non-roleplayers.

Using non-negative matrix factorization, we ran topic models for a non-roleplay guild and two roleplay guilds (see Table 1). The topic models reinforce the previous findings, demonstrating that the focus of roleplayers is largely on narrative content, whereas non-roleplayers are predominantly concerned with game objectives.

- Sentiment analysis to determine the extent to which roleplayers and non-roleplayers make use of emotive language.

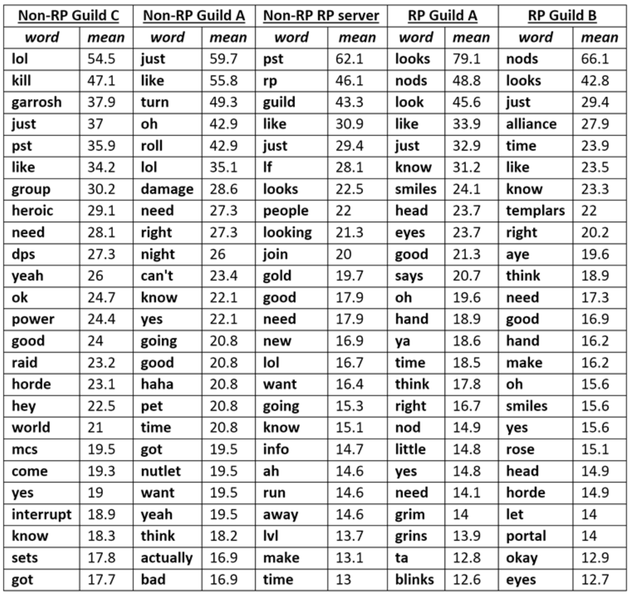
Sentiment analyses of the non-roleplay chatlogs reveals that the language of non-roleplayers, as would be expected, remains, from the perspective of sentimentality, consistent throughout. The results of the analysis of the roleplayers (see Fig. 3) are interesting in that the sentiment oscillates to a significant degree, demonstrating a high degree of verbosity in the language that they use.

Significance

While a difference between the play styles of roleplayers and non-roleplayers has often been assumed by researchers based on players' self-identification, our study confirms a real distinction between roleplay and non-roleplay in terms of language. We have shown quantitatively that the language of roleplay is more emotive, narrative, and verbose than that of non-roleplay. MMORPGs can provide a platform not only for interactive play, but also for interactive storytelling or group narrative construction.

Of particular significance is our finding that roleplayers frequently describe their avatars' body language. Through use of descriptors such as "looks", "nods", "stares", "glares", "grins", etc., roleplayers adapt to limitations in the control of their avatars by verbally recreating non-verbal cues. Analyzing the complex interplay of roleplayers' chat descriptions and the actual movements of their avatars can help us to better understand the challenges and potential of virtual forms of embodiment. It can also assist game designers in better accommodating the needs of a dedicated subpopulation of players.

Methodologically, our choice to seek chatlogs from roleplay and non-roleplay guilds, rather than rely solely on our own in-game chatlogs, is a strategy that can be of use to other researchers. Guilds in *World of Warcraft* frequently store—and, in the case of roleplay, even clean—their own chatlogs, which can provide a focused dataset.



Non-RP Guild C		Non-RP Guild A		Non-RP RP server		RP Guild A		RP Guild B	
word	*mean*	*word*	*mean*	*word*	*mean*	*word*	*mean*	*word*	*mean*
lol	54.5	just	59.7	pst	62.1	looks	79.1	nods	66.1
kill	47.1	like	55.8	rp	46.1	nods	48.8	looks	42.8
garrosh	37.9	turn	49.3	guild	43.3	look	45.6	just	29.4
just	37	oh	42.9	like	30.9	like	33.9	alliance	27.9
pst	35.9	roll	42.9	just	29.4	just	32.9	time	23.9
like	34.2	lol	35.1	lf	28.1	know	31.2	like	23.5
group	30.2	damage	28.6	looks	22.5	smiles	24.1	know	23.3
heroic	29.1	need	27.3	people	22	head	23.7	templars	22
need	28.1	right	27.3	looking	21.3	eyes	23.7	right	20.2
dps	27.3	night	26	join	20	good	21.3	aye	19.6
yeah	26	can't	23.4	gold	19.7	says	20.7	think	18.9
ok	24.7	know	22.1	good	17.9	oh	19.6	need	17.3
power	24.4	yes	22.1	need	17.9	hand	18.9	good	16.9
good	24	going	20.8	new	16.9	ya	18.6	hand	16.2
raid	23.2	good	20.8	lol	16.7	time	18.5	make	16.2
horde	23.1	haha	20.8	want	16.4	think	17.8	oh	15.6
hey	22.5	pet	20.8	going	15.3	right	16.7	smiles	15.6
world	21	time	20.8	know	15.1	nod	14.9	yes	15.6
mcs	19.5	got	19.5	info	14.7	little	14.8	rose	15.1
come	19.3	nutlet	19.5	ah	14.6	yes	14.8	head	14.9
yes	19	want	19.5	run	14.6	need	14.1	horde	14.9
interrupt	18.9	yeah	19.5	away	14.6	grim	14	let	14
know	18.3	think	18.2	lvl	13.7	grins	13.9	portal	14
sets	17.8	actually	16.9	make	13.1	ta	12.8	okay	12.9
got	17.7	bad	16.9	time	13	blinks	12.6	eyes	12.7

Figure 1

Overall, this research demonstrates the potential for increased intersection between games studies and critical DH methods. Our study provides the statistical evidence necessary to further extrapolate how roleplayers use language to create their own storylines, invent character personalities, and develop meaning in the context of a game's fictional world. In this paper, we will further detail the outcomes of our analyses, and offer a number of interpretations founded upon our results.

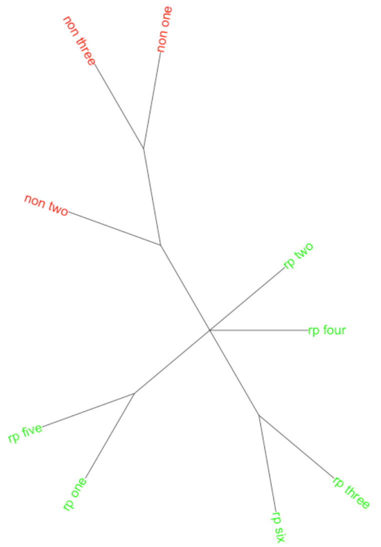


Figure 2

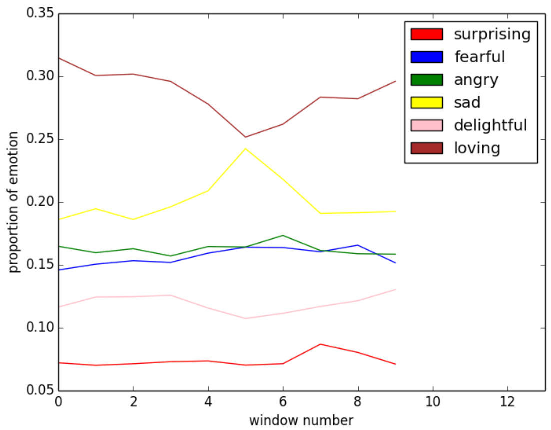


Figure 3

Top NMF topics in non-roleplay

Topic 0: kill mcs interrupt dps weapon switch power lol world corruption
Topic 1: power world horde whirling corruption lol true yes dps flows
Topic 2: rested feel longer learned blueprint watch frenzy goes thats just
Topic 3: pst lf guild wod just heroic looking soo man raid
Topic 4: loot corruption need help raid 100 tower arrogance free come
Topic 5: honor kill warchief fallen blood garrosh iron drown crush hold
Topic 6: lol like im yeah just ok think don good ll
Topic 7: noodles river fish harmonious think fresh hour caught curse island
Topic 8: roll weapons unfinished assembly begin line hey come gonna automated
Topic 9: group role lol queued just ok members raid selected initiated
Top NMF topics in roleplay (Guild A)
Topic 0: looks look nods eyes smiles blinks head just grins right
Topic 1: conversation 11 donnelly looks nods Sanctuary don
Topic 2: says nods da looks ah nod liene Sanctuary ta horde
Topic 3: ya like looks just know ta don oh drink good
Topic 4: Sanctuary ze nods peace justice looks nod mercy oh
Topic 5: looks smiles nods oh look ze good chuckles ve know
Topic 6: looks hand eyes just don look like know head doesn
Topic 7: looks look know just don nods like ve ll
Topic 8: rhenold te looks look oh smiles like smile don tat
Topic 9: nods Sanctuary looks like nod smiles look good don
Top NMF topics in roleplay (Guild B)
Topic 0: don right okay like ll think wolf oh um just
Topic 1: alliance yay vote nay yes terms rose nods just debt
Topic 2: portal makes gi strange gestures mog ha il team looks
Topic 3: nods mallory accused looks rann tribunal crimes silvergear questions little
Topic 4: nods ll need aye just guildB good time don know
Topic 5: morgan hand smiles bowl ashford firestar child light looks nods
Topic 6: horde iron time did looks nods know draenei speak foundry
Topic 7: looks nods draconic form head asea blinks eyes moment takes
Topic 8: alliance nods war GuildB looks king table aye order
Topic 9: girl glitter oh okay like grins time took kneels maybe

Table 1

Bibliography

Bainbridge, W. S. (2015). *The International Encyclopedia of Digital Communication and Society.* In R. Mansell, P. Hwa Ang, C. Steinfield, P. Ballon, S. Van der Graaf, A. Kerr, D. Kleine, (Eds.), John Wiley and Sons.

Barton, M. (2008). *Dungeons and desktops: The history of computer role-playing games.* Wellesley, MA: A K Peters, Ltd.

Corsini, R., Shaw, M. and Blake, R. (1961). *Role playing in business and industry.* New York: Free Press of Glencoe.

Notes

[1] Roleplay in videogames refers to users remaining "in-character" while playing the game. While roleplaying, users use performative communication, interacting in a manner suited to their character's personalities, and remaining constrained within the established limits of the game's backstory.

[2] A Massively Multiplayer Online Role-Playing Game, or

MMORPG, is a roleplaying videogame that takes place in a persistent Web-based world. This world is shared by a large player base interacting within a common environment for the purposes of socialising and playing the game in both an individual and collective manner.

Digital Annotation Tooling for Opera Performance Studies

Kevin Page
kevin.page@oerc.ox.ac.uk
University of Oxford, United Kingdom

Terhi Nurmikko-Fuller
terhi.nurmikko-fuller@oerc.ox.ac.uk
University of Oxford, United Kingdom

Carolin Rindfleisch
carolin.rindfleisch@music.ox.ac.uk
University of Oxford, United Kingdom

David Weigl
david.weigl@oerc.ox.ac.uk
University of Oxford, United Kingdom

Background

In opera and music theatre, the realisation of a work in performance differs significantly from the abstract concept that is captured in the score. The scenic interpretation, with its own characteristics and specific perspective on the work, is created afresh in every new staging – thus a performance and its experience cannot be determined from the score alone (Cook, 2013).

Comparing a mere three stagings of Richard Wagner's *Der Ring des Nibelungen* illustrates this point: the Bayreuth premiere in 1876, Boulez'/Cherau's *"centennial Ring"* in Bayreuth 1976, and the *Ring* Cycle in Birmingham in 2014 (which is the subject of our digital annotation capture in this paper). Not only are costumes and decorations entirely different, but also the setting and staging of the action, interpretation and presentation of the characters, visual and scenic aesthetics, and the perspective on and attitude towards the work.

For studies of the reception and perception of an operatic work, keeping a record of a particular performance – the characteristics of individual stagings – is an important requirement. This poses the question of how best to document the ephemeral phenomenon of an opera. Audio-visual recordings appear to provide an answer, but they are neither neither objective nor exhaustive; live annotation by a musicologist in the audience can thereby provide an additional or alternative resource.

Musical Score Annotation Kit

We developed the Musical Score Annotation Kit (MuSAK) to capture these musicological annotations, designing it to meet the requirements – and technical compromises – of operating in the environment and timescales of a live production in a working theatre, and providing an interface of sufficient responsiveness and adaptability to meet the needs of a working musicologist. The kit comprises:

- A touchscreen tablet device, running a bespoke web-based client through which each individual page of a score can be annotated and pages can be turned;
- A server based on the Union platform which gathers annotation keystrokes from the tablet client;
- A second "score following" page-turn annotation interface, capturing timings of the realisation (performance) of the material on a particular score page;
- A Livescribe Echo digital smart pen, used to take notes beyond those described in the annotation key (see below);
- Capture of audio and video, potentially recording both the staged performance and the musicologist using MuSAK.

Technical details of the MuSAK infrastructure are fully described elsewhere (Page et al., 2015). In this paper we report upon the utility of the toolkit: both in relation to the needs of the musicologist during the annotation event; and in relating the quantitative temporal data captured to the qualitative consideration of its potential musicological interpretation. Particularly with regard to the latter data-derived investigations, a key functionality of the kit for post-performance analysis is the temporal reconciling of the constituent digital media and annotations into a coherent metadata hyperstructure, enabling navigation of the information-dense digital captures (Nurmikko-Fuller et al., 2015).

Annotation workflow and enactment

The first full deployment of MuSAK captured annotations during a complete staging of Richard Wagner's *Ring*, performed on four nights over five days by the Mariinsky Opera at the Birmingham Hippodrome in November 2014 (Figure 1).

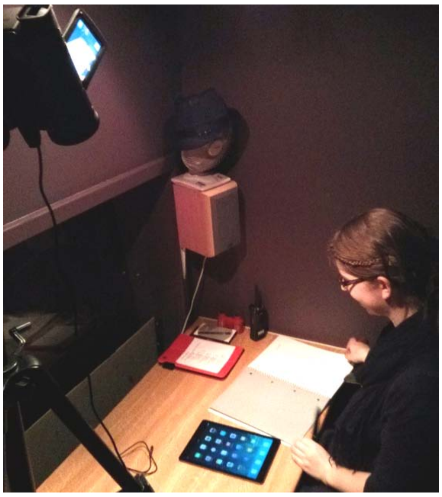


Figure 1: The musicologist and MuSAK in position from her viewpoint at the back of the stalls before curtains up

This multi-stage process is summarised in Figure 2. Digital images were generated from short piano scores of the operas obtained from IMSLP[1], then formatted and cleaned for viewing on the MuSAK tablet. The musicologist spent considerable time before the performance annotating printed copies of these scores which were subsequently re-digitised. These pre-performance notes highlighted musical elements determined directly from the score; these also served as "signposts" for potential points of interest which would be revisited during the live performance.

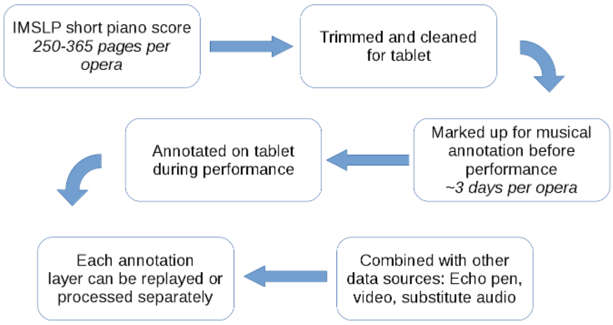

IMSLP short piano score
250-365 pages per opera
Trimmed and cleaned for tablet
Marked up for musical annotation before performance
~3 days per opera
Annotated on tablet during performance
Combined with other data sources: Echo pen, video, substitute audio
Each annotation layer can be replayed or processed separately

Figure 2: Annotation workflowscore_workflow.png

This necessitated the development of a symbolic key (Figure 3) which allowed annotations to be made at sufficient speed during the live performance, and which was trialled before deployment during co-development of the technical system. During the Birmingham staging the musicologist annotated the tablet-based score image directly using a stylus to draw symbols from the key (Figure 4), allowing for adaptation of the key between the nightly performances, based on experience during fieldwork.

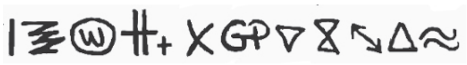


Figure 3: Examples of symbols from the musicologist's keykey_image.jpg

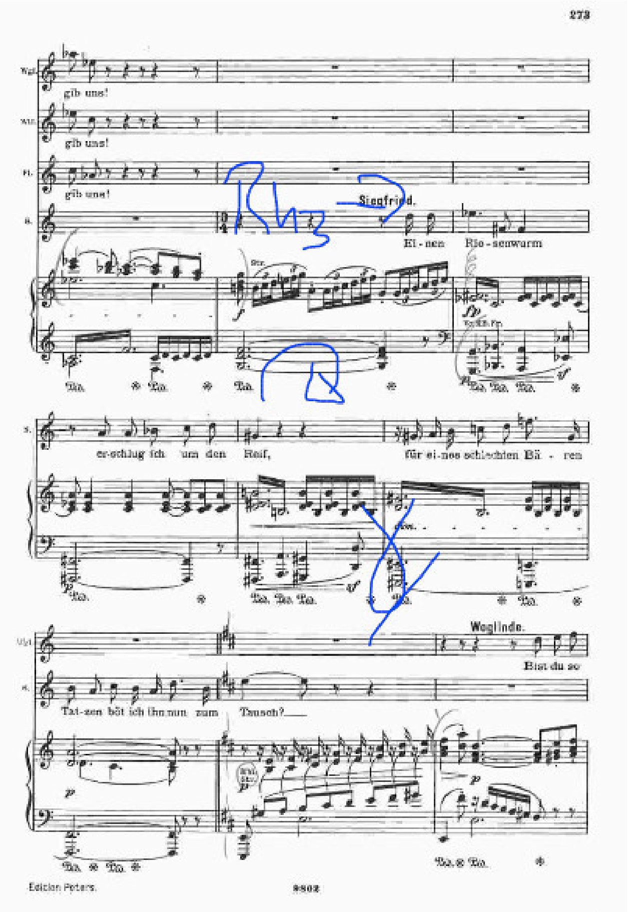


Figure 4: An annotated score pagescore_page.jpg

A significant volume of annotation and related data was gathered: 15 hours of video footage; over 100,000 tablet strokes encoding 8,216 annotations; 1,300 performance based page turns; 1,316 digital score images; and 104 pages of writing producing 13 hours of digital pen replay.

Toolkit Assessment

The usability attributes of MuSAK were assessed in interviews with the musicologist after the Birmingham performances according to learnability, efficiency, memorability, and satisfaction, as defined by Ferre (Ferre et al., 2001).

Learnability was evaluated through the musicologist's experience of acquiring the skills necessary to complete the annotation process. She found the system comparable to existing musicological annotation pragmatics, minimising the need for training:

> Using MuSAK [is] very similar to the process that I as a musicologist used to do regularly...I think it worked very well because [it repeated existing] processes [...and] fitted in with actions I was very well adapted to...the tools were very non-invasive.

On satisfactory functionality and performance, the musicologist felt "quite" able to *keep up with the pace of the* *opera*, although believed that the time necessary to make additional freehand annotations and cognitively process observations made the page turn annotations inaccurate:

> I was quite well able to keep up with the pace... I have been able to capture...a pretty good picture of...the profile of the performance... an important realisation is that making these scenic [and musical] annotations … which are particular about this performance... requires a lot of time to think and... process even if it is only [around] 10 seconds or 5 seconds...if we want to also capture a detailed and accurate time profile of the performance it would be necessary to have two people…

However, an analysis of the page turn data (below) indicates an ability to at least turn pages at a pace highly correlated with the performance. (MuSAK can also be configured to support multiple annotation devices and annotators simultaneously, although this was not done during the *Ring* capture.)

The musicologist reported an ability to *capture the idiosyncratic profile of each specific performance*, including deviations from the score or expectations based on it, as well as staging, lighting, and the behaviour of the actors:

> ...the performance details: are the musical details particular in any places, is something particular loud ...[or]... fast, are any mistakes made ... scenic specifics of the performance, what happens on stage... are people moving a lot, are they using a lot of gestures, are there significant lighting changes, are there changes of scenes...

MuSAK was described as *supportive of traditional annotation paradigms*, because new skills were not necessary for effective and efficient annotation using the touchpad screen and stylus:

> ...intuitive to use because it was mainly very similar to just using pen and paper which everyone who is concerned with analysing music is very used to... so it's very similar to the process that I as a musicologist used to do regularly anyway and ... it worked very well because it took up processes that were there anyway and tried to fit in with actions I was very well adapted to [...] very similar to [...] if you eventually get used to the touch screen which didn't take such a long time...

The additional affordances of a digital system were noted, including the automatic *capture of the temporal profile of the performance and the benefit of being able to easily make corrections:*

> ...they actually help because with using a pen and paper you wouldn't be able to undo things if you make mistakes, ...[nor]... be able to afterwards see the timings [...] with pen and paper doesn't give such an accurate profile of this particular performance [...] you're capturing the timings [of] not only the music but also the timings of the annotations because afterwards you can look at how dense are the annotations.

Data Interpretation

In addition to recording the musicologist's (subjective) annotations, MuSAK can be used to produce statistics about the acts of annotation to supplement our understanding of the technical system and our interpretation of the musicological context.

For example, Figure 5 shows comparative plots of what is, effectively, the rate of performance of each page (from the score-following annotations) compared to the musicologist, indicating an overall ability to "keep up" with the pace of the staging.

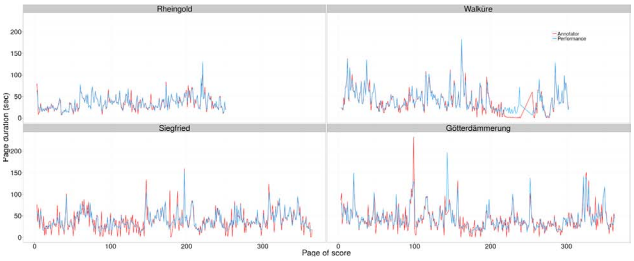


Figure 5: Comparative plots of performance and annotator

Figure 6 shows the rate of annotations over the course of the four operas, indicating what might be considered as a crude measure of "music information density".

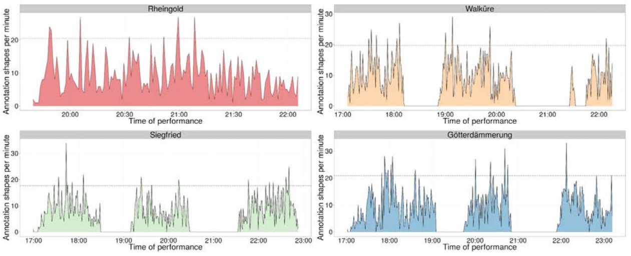


Figure 6: Rates of annotation over time during the Ring Cycle performance

Moments which pass a higher threshold of annotation rate have been identified to relate to significant music-dramatic instances in several places, which involve highly concentrated activity on stage (e.g. Hagen killing Gunther in Götterdämmerung).

Bibliography

Cook, N. (2013). *Beyond the Score: Music as Performance*. New York: Oxford University Press.

Ferre, X., Juristo, N., Windl, H. and L. Constantine, L. (2001). *Usability basics for software developers*, *IEEE Software*, **18**(1): 22–29.

Nurmikko-Fuller, T., Weigl, D. M. and Page, K. R.
On organising multimedia performance corpora for musicological study using Linked Data. *Proceedings of the 2nd International Workshop on Digital Libraries for Musicology.*Knoxville: ACM, pp. 25-28.

Page, K., Nurmikko-Fuller, T., Rindfleisch, C., Weigl, D. M., Lewis, R., Dreyfus, L. and De Roure, D. (2015). A Toolkit for Live Annotation of Opera Performance: Experiences Capturing Wagner's Ring Cycle. *Proceedings of the International Society for Music Information Retrieval 2015*. Malaga: ISMIR, pp. 211-17.

Notes

[1] The International Music Score Library Project (IMSLP)/Petrucci Music Library, http://imslp.org/

Bretez: Conjugaison du passé au futur

Mylène Pardoen
mylene.pardoen@wanadoo.fr
Institut des sciences de l'Homme de Lyon (FRE 3768) / CNRS

S'appuyer sur le sensible pour valoriser le patrimoine est une demande récente des musées. En effet, à une époque où le multimédia modifie chaque activité de découverte, agissant notamment sur la perception, il devient nécessaire d'intégrer cette dimension du sensible lors des présentations touchant au patrimonial. *Bretez* est un projet multimédia, associant des acteurs de sciences humaines et sociales et ceux de sciences de l'ingénieur. Il est conçu pour répondre à ces problématiques.

Le projet tire son nom du géographe (Louis Bretez) qui leva, en 1739, le plan pour le prévôt des marchands Turgot en utilisant l'axonométrie – une technique de projection révolutionnaire pour l'époque.

Le projet *Bretez* relève de la recherche empirique, exploratoire et d'une méthodologie expérimentale ayant pour objectif la mise au point d'un modèle. Sa conception s'appuie sur les techniques de virtualité. Comme tout modèle virtuel, *Bretez* n'incarne pas une vérité scientifique figée mais une proposition représentant l'état de la science à l'instant de sa création et qui est voué à évoluer. Dans notre cas, c'est un média (la maquette crée sur une plateforme de jeu) qui sert de départ pour le dialogue entre différents membres d'une communauté scientifique, mais aussi un public large.

Deux grands axes se dégagent dès le début des réflexions portant sur l'élaboration du projet *Bretez*. D'une part, comment intégrer la dimension sensible, d'autre part, comment répondre à cette demande de multimédia tout en respectant les besoins des musées.

L'archéologie du paysage sonore répond à la première de ces demandes spécifiques. L'écologie du sonore et son ouvrage référent *Le paysage sonore – le monde comme musique* de R. Murray Schafer (1977, réédition 2010), est la première étape d'une prise de conscience des spécificités de cet environnement qui forment un paysage sonore. Puis les études d'A. Corbin (*Les cloches de la terre* – 1994) et de J.-P. Gutton (*Bruits et sons dans notre histoire* – 2000) ouvrent un chemin vers les aspects sonores du passé. Mais ces études restent au stade du théorique. L'archéologie du paysage sonore permet de passer à l'étape expérimentale.

Que ce soit dans le cadre d'une restitution de paysages urbains, d'ambiances d'intérieur, ou autres, les problématiques restent identiques. La restitution ou création de visuels et de paysages sonores historiques pose, d'entrée, des questions fondamentales: pourquoi restituer, dans notre cas, ce Paris des Lumières? Peut-on « voir/entendre » ce passé? Comment le restituer, le faire entendre? Quelles sont les difficultés, les limites d'une telle restitution? Quels sont les matériaux? Où s'arrête l'acte du créateur, celui du chercheur? Comment dépasser le concept d'habillage sonore et aboutir la restitution d'un véritable paysage historique? Comment intégrer ces notions relevant du sensible, les porter au public? Pour quel public?

La demande muséographique (dans notre cas) place le concepteur aux confins de la composition, de la musicologie et de l'histoire. Un « simple » habillage sonore (*sound design*) ne peut, en effet, satisfaire cette demande très spécifique qui entre, également, dans les cadres de la sauvegarde du patrimoine immatériel.

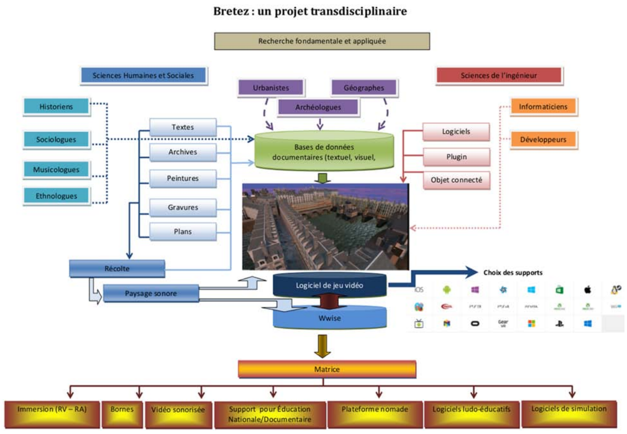
En effet, pour le musicologue, pour le metteur en sons, ce travail est inhabituel. Les matériaux ordinaires sont *quasi* inexistants: pas d'enregistrements disponibles pour les années antérieures à 1870. Entendre le passé implique donc qu'il faut en chercher les traces dans l'ensemble du corpus littéraire (livres, journaux, etc.) ou graphiques. Puis, tenter de le pister dans le temps d'aujourd'hui afin d'en faire une captation. Il faut également solliciter l'imaginaire lors de la restitution pour mieux servir l'Histoire.

Pour le visiteur, l'ambiance sonore ainsi (re)créée doit également exciter son imaginaire pour l'aider à mieux percevoir, recontextualiser des situations historiques, tout en préservant la réalité historique et non jouer sur les émotions. Trouver le juste milieu dans la sollicitation de cet imaginaire pour mieux servir le patrimoine et le mettre en valeur sera le fil conducteur de cette première partie.

D'autre part, ainsi que nous l'avons mentionné plus haut, dès nos premières observations sur et autour du Projet *Bretez*, il est apparu que de nombreuses contraintes techniques étaient présentes, notamment des contraintes matérielles et logicielles. Parmi celles-ci, l'obsolescence présente un réel frein pour les musées car elle représente un coût financier important. Il devint dont évident que mettre au service des musées, des historiens ou des archivistes une représentation visuelle, spatiale et audio qui respecte et témoigne d'une réalité du quotidien était un enjeu majeur, ceci afin de faciliter sa communication à un public de plus en plus friand de multimédia. Il était également incontestable que nous pouvions y répondre grâce, notamment, à l'utilisation d'un moteur de jeu.

Ce détournement d'utilisation de logiciel de conception de jeux vidéo soulève, à son niveau, de nombreux problèmes épistémologiques et éthiques. Comment développer une telle plateforme à des fins pédagogiques et scientifiques? Comment utiliser l'outil? Quelle stratégie de conception et de scénarisation? Quelles en sont les avantages, mais aussi les limites? Pour quel public? Ces réflexions structureront cette seconde partie.

La présentation de mes travaux s'appuie sur la maquette qui sert de matrice à la recherche et s'attardera plus particulièrement sur l'élaboration d'une maquette historique sonore, en s'articulant autour de cette réflexion: comment voir et ouïr le passé pour réjouir nos sens sans trahir l'Histoire?



Bretez est un projet restitution de Paris au XVIIIe siècle. La maquette virtuelle, résultat d'une coopération scientifique, associe des équipes de sciences humaines et celles des sciences de l'ingénieur (Humanités Numériques). Sa destination première est orientée muséographique. Son objectif, la valorisation patrimoniale par sa restitution numérique tridimensionnelle et sonore spatialisée, se dénote par une nouvelle approche de la restitution du passé en 5 D (la combinaison du visuel – la 3D –, le déplacement et la dimension sensoriel – le sonore). Fruit d'un travail collectif, *Bretez* présente une très forte dimension sonore qui rend le passé disponible et tangible pour un très large public. Il met ainsi en valeur une vision méconnue et scientifiquement valide de la ville lumière.

À terme, ce projet vise la conception d'une plateforme générique pour la génération procédurale de villes virtuelles, tant sonore que visuelle, qui permet de s'affranchir des contraintes technologiques liées à la diversité des supports et leur obsolescence. Projet innovant, ses spécificités porte sur:

- L'inter et la transdisciplinarité (association étroite entre les Sciences Humaines et Sociales et les Sciences de l'ingénieur);
- La valorisation du patrimoine par le sensible;
- La construction autour du son et une élaboration concomitante visuel/son.
- La création d'un nouveau concept: la 5D.

Un tel projet permet d'élargir au maximum sa sphère de rayonnement. Pensé pour la muséographie, il s'est orienté vers la recherche sur demande d'équipes de chercheurs historiens ou archéologues (notamment LaDéHiS/EHESS). Il n'entre pas dans la catégorie jeu ou purement artistique car répond à des critères scientifiques. Toutes les sources sont primaires (archives de police, notariales, témoignages d'époque…). Nous assumons les « vides » : S'il n'y a pas de sources (ou ne pouvant être validées par la communauté scientifique), nous n'inventons pas – tant sur le plan visuel que celui des ambiances sonores. Chaque information est recoupée et validée par des historiens, sociologues... Le tout forme autant de contraintes à tout acte compositionnel « traditionnel ». Les questions de temporalités, au sein du logiciel, sont réétudiées afin de répondre aux critères historiques et scientifiques et non ludiques.

L'élaboration de la maquette initiale sur une plateforme de jeu vidéo est un véritable atout. Ce type de logiciel permet de répondre à des demandes diverses et ouvre la porte à l'élaboration de toute sorte de produits finalisés : ceux touchant à la réalité virtuelle, à la réalité augmentée, à la création de supports immersif ou non, à ceux nomades (pour les principaux)... Le projet entre en phase de prototypage. Les premiers essais et retours sont programmés à partir de juin 2016. Dans un premier temps, il est prévu la mise en test *in situ*, Place du Châtelet à Paris, d'une application en réalité augmentée en partenariat avec Labo-M (chef de projet: Pierre Girard) et le musée Carnavalet.

Ce projet soutenu et accompagné par la SATT PULSALYS pour une mise en chantier au 01/01/2016 https://sites.google.com/site/louisbretez/home

Bibliography

Aubrun J., Bruant C., Kendrick L., Lavandier C. and Simonnot N. (2015). *Silences et bruits du Moyen-Âge à nos jours. Perceptions, identités sonores et patrimonialisation*, L'Harmattan, Paris.

Beck R., Kramp U. and Retaillaud E. (2013). *Les cinq sens dans la ville du Moyen-Âge à nos jours*, Presses Universitaires François Rabelais, coll. Villes et territoires, Tours.

Gauthier L. et Traversier M. (sous la direction de) (2008). *Mélodies urbaines – la musique dans les villes d'Europe (XVIe-XIXe siècles)*, PUPS, Paris.

Gutton J.–P. (2000). *Bruits et sons dans notre histoire*, PUF, Paris.

Mercier L. S. (de 1782 à 1788//1979). *Tableau de Paris*, Amsterdam (en douze volumes)/Réimpression, Slatkine Reprints, (6 volumes), Genève.

Murray S. R. (2010). *Le paysage sonore – Le monde comme musique*, 4e édition, Wildproject, France.

Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish

Adam Tomasz Pawłowski
adam.pawlowski@uwr.edu.pl
University of Wrocław, Poland

Chronological corpus – definition

The notion of chronological corpus has appeared so far only in publications of limited circulation (cf. Pawłowski, 2006), whereas it is practically absent from mainstream literature on corpus linguistics and digital humanities. This is somewhat surprising given that the research of so called "lexical series" has been conducted in the past (cf. Brunet, 1981; Salem, 1987) and that Google has been for some years developing the tools *Ngram Viewer* and *Google Trends*, which allow sequential analysis of Google Books resources and of users' Internet queries (cf. Michel et al. 2011).

In order to understand what chronological corpus is one must grasp the basic distinction between synchrony and diachrony at the heart of structural linguistics. Synchrony is a study of language at a certain moment in time, where "moment" can last even several decades, provided there are no apparent changes in grammar, vocabulary and pronunciation. As for diachrony, it exposes evolution of language taking place in the process of its development, usually over long periods, spanning even several centuries. However, corpus research supported by NLP tools enables a far more flexible approach to the time variable, since text samples consistent in terms of their typography, spelling and grammar can be annotated with their exact publication dates. Scientific description then focusses on the dynamics of the frequency change of specific lexemes (or other segments) in time, rather than on the evolution of word forms beginning with a hypothetical proto–language until the present state.

An important issue to be resolved is that of naming such an approach, whereby the chosen term would instate it in the domain of corpus linguistics. While social sciences and psychology use the term of "longitudinal" research, its equivalent in linguistics could be 'microdiachrony'. Having said that, the term's association with traditional diachrony and history of language could prove misleading. This is why it is more advisable to speak of **chronological analysis** and **chronological corpora**, where, similarly to diachrony, both terms contain a reference to time – gr. *chronos*.

Consequently, **chronological corpus should be understood as a sequence of text samples consistent in terms of their spelling and grammar, corresponding to subsequent points on the axis of time** (e.g. weeks, months etc.). Such a corpus would make it possible to explore the dynamics of the change of lexeme frequencies in successive periods using the method of time series analysis (cf. Gottman, 1981; Cryer, 1981; Pawłowski, 1998, 2001). A chronological corpus would differ from a diachronic one in that in the former texts are evenly spread in time and word forms remain unchanged, while in the latter it is the opposite: word forms must evolve to become object of interest and time spans between measurements may be of any length. It needs to be emphasised that the annotation of the time variable may be ignored (if such is the premise of the conducted research); a chronological corpus would then be treated as a synchronic one.

Chronological corpora

There exist corpora that are intrinsically arranged with regard to the time variable, e.g. literary outputs of some authors presented as Shakespearean Corpus, Corpus Thomisticum or Corpus Platonicum. If the date of creation of every piece of data in the corpus is known, the development of the text's stylostatistical properties over time can be established. If text chronology remains partly unknown, stylometric research can help to put undated works in the correct order (cf. Lutosławski, 1897).

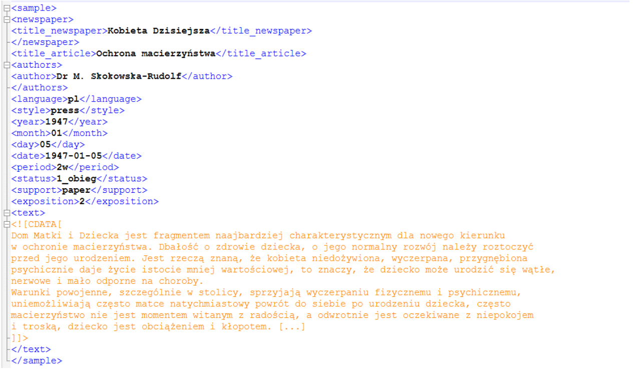
But there is also another set of text corpora that seem to be naturally predestined to sequential analysis. Contrary to other genres or functional styles, **press and media texts** necessarily incorporate the date or even the time of their creation. This is of course little surprising, since media must comment upon actual events. Publication dates in the headers or footers of printed newspapers may be thus regarded as "time markup" that is invaluable as a Gutenberg era contribution to the sequential corpus research in the digital universe.

ChronoPress – features and origin

At present ChronoPress corpus contains texts representing Polish press of the post-war totalitarian period (1945-1954). Future extensions are planned to cover the entire period of the country's sovereignty (1918–2018). As the flow of press information was abundant even in the early post-war period, representative method of sampling has been applied. It has also been assumed that the minimum time spans, visible from the user level, are limited to subsequent months (the volume of text corresponding to weeks was too small to guarantee a sufficient level of representativeness). Apart from that it should be pointed out that the pace of public discourse in the times of printed press was slower than it has been in the digital world, so that months as time units can be regarded as sufficient to trace relevant events or processes.

In order to obtain reliable values of lexeme frequencies the number of words per month has been established as ca 120.000-140.000. Since sample size has been fixed to ca 250-300 words of continuous text, every month is represented by 480 samples and every year by 5760 samples (ca 1,5 million words per year). Equal numbers of words per month are important to guarantee that time series and other parameters generated from ChronoPress remain statistically unbiased with regard to the sample size. Selected newspapers represent a relatively broad spectrum of public post-war discourse of the totalitarian state. Particular titles to be included in the corpus have been selected according to two criteria: newspaper circulation (only titles with the highest values and national range have been considered) and target groups, as defined by the ideology of the time (first of all the working class, peasantry, army, youth, and, to a lesser extent, "intelligentsia"). The corpus is freely available through an interactive user interface.

Data have been thoroughly sampled from Polish dailies and weeklies. Preparatory stages of analysis consist of: sample selection, scanning, OCR, markup with metainformation (XML scheme) and text curation. The corpus has been annotated morphosyntactically using Morfeusz tool which had been designed for the Polish National Corpus. Metadata (Fig.1) include newspapers' titles, articles' titles, authors' names, exposition, and data support.



```xml
<sample>
<newspaper>
<title_newspaper>Kobieta Dzisiejsza</title_newspaper>
</newspaper>
<title_article>Ochrona macierzyństwa</title_article>
<authors>
<author>Dr M. Skokowska-Rudolf</author>
</authors>
<language>pl</language>
<style>press</style>
<year>1947</year>
<month>01</month>
<day>05</day>
<date>1947-01-05</date>
<period>2w</period>
<status>1_obieg</status>
<support>paper</support>
<exposition>2</exposition>
<text>
<![CDATA[
Dom Matki i Dziecka jest fragmentem naajbardziej charakterystycznym dla nowego kierunku
w ochronie macierzyństwa. Dbałość o zdrowie dziecka, o jego normalny rozwój należy roztoczyć
przed jego urodzeniem. Jest rzeczą znaną, że kobieta niedożywiona, wyczerpana, przygnębiona
psychicznie daje życie istocie mniej wartościowej, to znaczy, że dziecko może urodzić się wątłe,
nerwowe i mało odporne na choroby.
Warunki powojenne, szczególnie w stolicy, sprzyjają wyczerpaniu fizycznemu i psychicznemu,
uniemożliwiają często matce natychmiastowy powrót do siebie po urodzeniu dziecka, często
macierzyństwo nie jest momentem witanym z radością, a odwrotnie jest oczekiwane z niepokojem
i troską, dziecko jest obciążeniem i kłopotem. [...]
]]>
</text>
</sample>
```

Fig. 1 Example of sample annotation with metadata

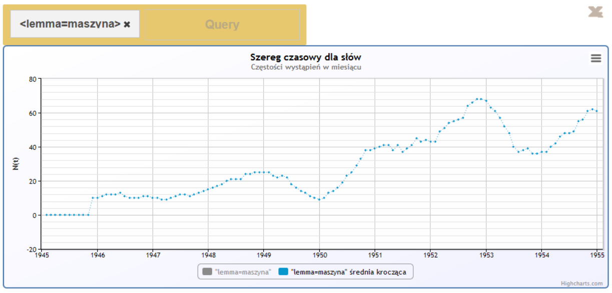


Fig. 2 Moving average of the frequency of the lexeme *machine* (maszyna). It supports the view that despite the systemic inefficiency of communist economy technological progress was promoted in post-war Poland.

ChronoPress offers a variety of online linguistic and exploratory tools, such as concordances, frequency lists, word profiles and word maps. Its specific feature is, however, the possibility to generate time series. The analysis of lexeme frequency values displayed on an axis of time allows users to discover and explore the dynamics of events and phenomena represented in daily press over long periods of time.

Goals of the presentation

ChronoPress allows to conduct a wide range of research in linguistics, contemporary history, cultural anthropology, media studies and communication science. The perspective adopted here is oriented towards the analysis of dynamic processes occurring in time rather than towards static states of language reflecting extralinguistic reality. The goal of the presentation is to provide an overview of the available exploratory techniques for chronological corpora, highlighting their explanatory power and limitations. In particular the following issues will be addressed and exemplified with empirical evidence from the ChronoPress corpus:

Manual and automated trend detection, modelling and interpretation

Long-term change in frequency of selected lexemes will be analysed as the expression of dominating topics of interest in public discourse, but also of social change imperceptible from the point of view on an average. Illustration material will include salient and informative examples generated from the ChronoPress database (*inter alia* lexemes from the semantic fields *war*, *health*, *technology*). Since trend analysis requires some conceptualisation of time, various approaches to time will be presented and discussed (linear or circular, anthropological, political and astronomical).

Detection of events

Events are understood here as sudden changes, such as natural or technological catastrophes, wars, accidents etc. They manifest themselves as sudden rises in the lexeme frequency. This task will be illustrated with some examples generated from the ChronoPress corpus and from other online sources.

Testing the efficiency of statistical tools

The utility of the autocorrelation function (for univariate series) and of the cross-correlation function (for multivariate series) in the analysis of chronological corpora will be verified.

Comparison of ChronoPress and online resources

Informational potential and functionalities of the ChronoPress corpus will be compared with other online research tools, such as Google Ngram Viewer, Hansard Corpus and Polish National Corpus.

Word profiles analyses

Collocations and semantic profiles of selected lexemes in subsequent years will be generated (using Log Likelihood Ratio) and compared.

Automatic detection and annotation of named entities

Proper names and other named entities in Polish can be detected and annotated using Liner 2. Spatial distribution of toponyms identified by Liner 2 will be automatically displayed on a map and further analysed. Both tools have been created in the framework of the Clarin-PL project (http://clarin-pl.eu/en/home-page/).

Bibliography

Brunet, E. (1981). *Le vocabulaire français. De 1789 à nos jours.* Paris–Genève: Slatkine, Champion.

Cryer, J. (1986). *Time series analysis*. Boston: Duxbury Press.

Gottman, J. M. (1981). Time-series analysis: a comprehensive introduction for social scientists. Cambridge, London etc.: Cambridge University Press.

Lutosławski, W. (1897). *The origin and growth of Plato's logic*. London, New York, Bombay: Longmans, Green and Co.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 14 (2011), Vol. 331: 176–182.

Pawłowski, A. (1998). *Séries temporelles en linguistique. Avec application à l'attribution de textes: Romain Gary et Émile Ajar*. Paris, Genève: Champion-Slatkine.

Pawłowski, A. (2001). *Metody kwantytatywne w sekwencyjnej analizie tekstu* [Quantitative methods in sequential text analysis]. Warszawa: Uniwersytet Warszawski, Katedra Lingwistyki Formalnej.

Pawłowski, A. (2006). Chronological analysis of textual data from the Wrocław Corpus of Polish. *Poznań Studies in Contemporary Linguistics* (PSiCL) 41, 2006: 9-29.

Salem, A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris: Klincksieck.

Social Networks and Archival Context: People and Cultural Heritage

Daniel Pitti
dpitti@Virginia.edu
Institute for Advanced Technology in the Humanities, University of Virginia, United States of America

Worthy Martin
wnm@eservices.virginia.edu
Institute for Advanced Technology in the Humanities, University of Virginia, United States of America

Overview

Social Networks and Archival Context (SNAC), initiated in 2010 as a R&D project, is now being transformed into an international cooperative. SNAC's original research objective was to demonstrate that descriptions of people, embedded in the descriptions of historical records that document their lives, could be extracted and used to reveal the social networks within which their lives were lived *and* provide integrated access to geographically dispersed historical records. SNAC's early success led to plans to establish a sustainable international cooperative to maintain and expand these descriptions of people. The long-term technological objective is a platform to support a continuously expanding, curated corpus of reliable biographical descriptions of people linked to, and providing contextual understanding of, the historical records that function as primary evidence for understanding their lives and work. The SNAC Cooperative will benefit librarians, archivists, and researchers, and will provide traditional historical researchers integrated access to distributed historical records and the social contexts within which the records were created and used. It will provide prosopographic researchers with methods for reconciling and establishing reliable social networks, and will enable archivists and librarians to share descriptive data while also making descriptions more effective.

The Archival Description Source Data

Archival description source data encompasses both descriptions of historical records as well as authority data for corporate bodies, persons, and families documented in historical records. OCLC WorldCat, sixteen archival consortia (representing hundreds of individual repositories), over thirty repositories, and two digital humanities research projects contributed their source data to SNAC. The holdings of over 4000 repositories are represented:

- 190,000 finding aids, contributed by fifteen consortia and over thirty repositories in the U.S., the ArchivesHub in the U.K., and the Bibliothèque nationale de France (Catalogue Collectif de France (CCFr) and BnF archives et manuscrits)

- 2.25 million OCLC WorldCat archival descriptions
- 400,000 authority records contributed by NARA (93,051), the British Library (297,731) the Smithsonian Institution Archives (2,083), the New York State Archives (258); and the Archives nationales, France (2,350)
- 30,000 correspondent names from the Joseph Henry Papers Project, Smithsonian Institution Archives
- 2,332 correspondents from The Walt Whitman Archive
- 1,200 names associated with the Chaco Research Archive

Data Processing

During SNAC's R&D phase (2010-2015), this source data was processed in three distinct steps.

- Biographical and historical data was extracted or migrated from existing archival descriptions and assembled into standardized descriptions that identify and document organizations, persons, and families based on an international archival communication standard, Encoded Archival Context – Corporate Bodies, Persons, and Families (EAC-CPF). Each EAC-CPF identity description includes the description of the entity as such (names, life dates, biographical information, etc.), links to descriptions of the historical records from which the data was derived, and links to other identities found in the same source. These links provide the foundation for assembling a vast social-document network or graph.
- The EAC-CPF identity descriptions were matched (identity reconciliation) against one another and against descriptions in the Virtual International Authority File, combining records that identify the same entity, to produce a set of unique EAC-CPF records.
- We developed a prototype access system, based on Extensible Text Framework (XTF), open source software from the California Digital Library. It has three major functional components: 1) display of the EAC-CPF records; 2) sophisticated searching and exploration of the EAC-CPF records; and 3) exposing the data to enable third-parties to access and use it in other applications.

Extracted or Migrated Data

The first step resulted in 6,719,064 Encoded Archival Context – Corporate Bodies, Persons, Families (EAC-CPF: an archival encoding standard hosted by the Society of American Archivists and developed in collaboration with the international archival community).

- 4,653,365 Persons
- 1,868,448 Corporate Bodies
- 197,251 Families

Merged Data

After performing identity resolution processing (match and merging), we had:

- 3,741,262 EAC-CPF records
- 2,466,425 persons1,077,588 corporate bodies197,249 families[1]
- 7,966,737 links between the 3,741,262 persons, corporate bodies, and families
- 15,031,209 links to 2,079,504 unique resource descriptions

Prototype History Research Tool

The prototype history research tool (http://socialarchive.iath.virginia.edu/snac/search) allows researchers to find persons, organizations, and families; to read biographic information about them; to explore the social networks within which they existed; to locate historical records that document their lives, related resources, and external links associated with that name. Associated links are provided for ArchivesGrid and Digital Public Library of America, as well as "sameAs" links to Wikipedia, VIAF, WorldCat Identities, and others.

Significance for Researchers

Researchers have welcomed SNAC for its research economies: SNAC's History Research Tool provides integrated access to distributed primary (archival) and secondary (published) resources, eliminating or at least substantially ameliorating the need to track down resources in multiple archival catalogs. Painstakingly locating these resources is a labor-intensive, time-consuming activity in the current research environment, with successful discovery and assembling of the data highly dependent on persistence and serendipity. Indeed it is likely that some of the information found in the SNAC records might never be discovered using current methods. SNAC also makes explicit what has been, at best, implicit in archival description: the social-professional-intellectual networks within which the lives and work of the people documented in historical resources took place. It exposes the vast global social-document network that connects the past to the present. Ed Ayers, President of the University of Richmond and a Civil War historian, wrote that:

> SNAC promises to change the way history is imagined and written! For all that the digital revolution has revolutionized, the heart of research lies within the primary record embedded in archives large and small. The pioneering work of SNAC will unlock that record, revealing connections and patterns invisible to us now.

Alan Liu, Professor of English, University of California, Santa Barbara and Director of Research Oriented Social Environment (RoSE), describes SNAC's potential:

> SNAC employs state-of-the-art computational techniques to do three things very well: 1) unlock information originally recorded for specific purposes in library and other archival finding aids to make them usable in new contexts; 2) connect widely-distributed information of this sort from around the world; and 3) marry the "library" or "archive" model of knowledge to a whole other model of social networks that both humanizes our understanding of the way knowledge emerges from communities of knowledge creators and seekers, and speaks powerfully to today's "social network" generation.

Significance for Prosopographical Research

SNAC is building a humanities resource that benefits humanities researchers, but ongoing development and refinement of identity reconciliation techniques are of further benefit to humanists engaged in prosopographical research. Names alone are weak identifiers: multiple people can have the same name and one person may have multiple names. A number of factors influence our ability to reliably identify people. Indeed, the larger the domain from which names are drawn, the higher the likelihood that a name is shared by several people.

Though each step in the processing described above presents intellectual and technical challenges, the most challenging is identity reconciliation. A fundamental human activity in the development of knowledge involves the identification of a unique "real world" entity (e.g., a person or book) and recording facts that, when taken together, uniquely distinguish that entity. Establishing the identity of a person, for example, involves examining available evidence, including the existing knowledge base, and recording facts associated with him or her (such as names, dates and places of birth and death, occupation, etc.). This is an ongoing, cumulative activity that both leverages existing established identities and establishes new identities. Identity reconciliation is the process by which an encountered identity is compared against established identities, and if not matched, is itself contributed to the established base of identities. The networked computing environment presents opportunities for using algorithm-based inference methods to compare newly encountered entities with established identities to determine the probability that a new entity represents the same person or thing as an established identity. This ongoing expansion of the base of reliable identities is an interplay of human research, knowledge recording, and computational methods.

Transforming SNAC into an International Cooperative

It became clear early on that the biographical data extracted and assembled from archival resource description constituted a valuable independent resource that could (and should) be maintained and further developed cooperatively. Development of a cooperative began back in 2011 and it recently entered its pilot phase with a group of fourteen inaugural institutional members that support the potential benefits of aggregated description and access demonstrated to date in SNAC, and, further, embrace the idea that the resources amassed should be cooperatively built and maintained in order to fully realize these benefits. The initial members represent research archives, libraries, museums (art and natural history), government archives, and institutional archives. The U.S. National Archives and Records Administration (NARA) serves as the secretariat for the Cooperative, while the Institute for Advanced Technology in the Humanities (IATH), University of Virginia, hosts the technological infrastructure. SNAC is led by IATH, working collaboratively with NARA, the California Digital Library, and the iSchool at UC Berkeley. The National Endowment for the Humanities (2010-2012), the Institute for Museum and Library Services (2011-2013), and the Andrew W. Mellon Foundation (2012-2017) have provided funding for SNAC.

Notes

[1] Because family names, as traditionally formed, lack sufficient qualifying information and thus commonly result in false positives, no matching was done against family names. In the final production, two family names were rejected as malformed.

Early English Books in Context: Towards a History of the Technological Humanities

Daniel James Powell
djpowell@uvic.ca
University of Victoria, Canada

Writing in *Literary and Linguistic Computing*, Julianne Nyhan et al argue that "without a better understanding—a more appropriate term might be "body of interpretations"—of the near and distant history of computing in the humanities, we are condemned to repeat the revolutionary trope *ad infinitum*." (Nyhan, Flinn, and Welsh, 2013). Willard McCarty, amplifying this, writes that "rather than hypnotizing ourselves with supposedly unprecedented marvels, we must learn to see computing in its historical and social contexts, and so be able to ground modelling in something larger than itself. For computing to be *of* the humanities as well as *in* them, we must get beyond catalogues, chronologies, and heroic firsts to a genuine history. There are none yet." (McCarty, 2008).

Susan Hockey wrote in *Companion to Digital*

Humanities that "humanities computing has a very well-known beginning," by which she means the decades-long collaboration between Father Roberto Busa and IBM to create a concordance of the work of Thomas Aquinas (Hockey, 2004). This is the heroic first, which Busa, writing in the forward of that same volume, summed up with admirable brevity: "During the World War II, between 1941 and 1946, I began to look for machines for the automation of the linguistic analysis of written texts. I found them, in 1949, at IBM in New York City."– This narrative is familiar to many of those working in digital humanities today, and has become openly accepted as the standard historical background for, first, humanities computing and, subsequently, the digital humanities writ large.

This presentation aims to upset this easy narrative by re-situating the history of one type of digital humanities project—Early English Books Online—as one chapter in an overall history of a technological humanities. That history—the history of a technological humanities—the story of how academics have deployed technology to better understand human creations, especially in textual form—or to understand and explore texts—did not begin in 1949. The creation of digital humanities, radiating outward from those early years, is surely part of the larger story of how technology and text have come together and drifted apart over many centuries. I claim that there is a great deal more continuity in the apparatuses, in the knowledge infrastructure of the humanities than we put forward in our "official" histories. In our search for a neat disciplinary history, we elide technology as a whole with the digital electronic computer. Busa's project likely does represent the beginning of one type of computational textual processing. It bears remembering, however, that his goal was to create a concordance, a type of reference tool and interface in existence in Western Europe since at least the 13th century. What is the history of *this* type of textual processing in the intervening six hundred years?

Instead of a history of tools for textual work beginning with the rise of humanities computing and moving forward to the present day, I hope to juxtapose a different narrative, one that troubles the rhetoric of a textual digital humanities that arises from a clear break with what came before. I hope to, perhaps polemically, test the boundaries of histories of digital humanities by considering an equally technologically sophisticated pre-digital humanities. Such a reframing opens many avenues of inquiry, including a consideration of Linked Open Data in the context of cooperative cataloguing practices from the 19th & 20th centuries, or contemporary textual analysis tools such as Voyant alongside the imposing machinery of an electro-mechanical Hinman Collator. This presentation, however, will highlight particularly those technologies of textual reproduction developed prior to the oft-quoted originary moment of 1949. Drawing on the history of Early English Books Online (EEBO), I argue that while a *computational* humanities may indeed be limited to the last half-century, the *technological* humanities—in both materialist and cultural senses—have a much longer history.

ProQuest introduces the resource on their front page:

> From the first book printed in English by William Caxton, through the age of Spenser and Shakespeare and the tumult of the English Civil War, Early English Books Online (EEBO) will contain over 125,000 titles listed in Pollard and Redgrave's Short-Title Catalogue (1475-1640), Wing's Short-Title Catalogue (1641-1700), the Thomason Tracts (1640-1661), and the Early English Tract Supplement - all in full digital facsimile from the Early English Books microfilm collection.[1]

In practice, this means that users are able to view the metadata for a given text; view page images of the original, early modern books in TIFF or PDF format; and, where available, view a full text transcription of the volume that are derived from the EEBO - Text Creation Partnership. Efforts to microfilm early English books began in 1931, intensified as World War II loomed, and continue today. Digital images of these microfilmed documents were made (and are still being made) available online first in 1998. The printed *Short Title Catalogue* (itself published in 1926) has determined what objects were photographed and, subsequently, scanned and put online–(EEBO).

The history of EEBO crosses multiple media, was directly impacted by global war, involves private companies and public universities, and is both analog and digital. To bracket EEBO (or EEBO-TCP) as a only a digital project impoverishes our understanding of how digital technologies have impacted the reproduction, preservation, and use of texts in humanistic scholarship. To write the full history of the early English books project, Early English Books Online, the Early English Books Online Text Creation Partnership is to engage in an act of disciplinary archaeology, one that forces digital humanists to grapple with the pre-digital origins and ideologies that inflect contemporary digital resources undergirding scholarship. EEBO is a microcosm through which one body of interpretations of digital humanities might be seen.

As much as it is a history, this presentation is also engaged in answering claims by Alan Liu and Tara McPherson, amongst others, that digital humanities has chosen disciplinarily to disengage from socio-critical questions.–[2] Thinking through the history of EEBO is one way to approach the digital humanities as a discipline tied to war-driven technological development; the uneasy relationships between private-sector providers and our shared cultural heritage; or the many varieties of labour that are imbricated within the knowledge infrastructures humanists use day in and day out.–[3] Blending media analysis, historical perspectives, and in-depth knowledge of humanities research tools, I hope to question the boundaries of what we consider digital humanities to be, how we write our histories, and how we move forward.

Bibliography

About EEBO.*Early English Books Online*. <http://eebo.chadwyck.com/marketing/about.htm>

Hockey, S. (2004). The History of Humanities Computing. *A Companion to Digital Humanities*. (Eds.) Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell.

McCarty, W. (2008). What's going on? *Literary and Linguistic Computing*, **23**(3): 253-61, doi: 10.1093/llc/fqn014.

Nyhan, J., Flinn, A. and Welsh, A. (2013). Oral History and the Hidden: Histories project: towards histories of computing in the humanities. *Literary and Linguistic Computing*, doi: 10.1093/llc/fqt044.

Svensson, P. (2009). Humanities Computing as Digital Humanities. *Digital Humanities Quarterly*, **3**(3). http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html.

Notes

[1] This text was current as of summer 2015 and is available in cached form. Since that time, Proquest has altered the front page description of EEBO to the following: Early English Books Online (EEBO) contains digital facsimile page images of virtually every work printed in England, Ireland, Scotland, Wales and British North America and works in English printed elsewhere from 1473-1700 - from the first book printed in English by William Caxton, through the age of Spenser and Shakespeare and the tumult of the English Civil War. Strangely, this newer version eliminates reference to the Early English Books microfilm Collection, as well as collapsing a number of distinct early modern collections of content into what might be called the EEBO brand. See the current version of <http://eebo.chadwyck.com/home> and a cached version <https://web.archive.org/web/20150905141338/http://eebo.chadwyck.com/home> from September 2015.

[2] For work by Liu on this topic, see "Where is the Cultural Criticism in the Digital Humanites?," published in *Debates in the Digital Humaniteies*, ed Matthew K. Gold <http://dhdebates.gc.cuny.edu/debates/text/20>. For McPherson work on UNIX and ideologies of race, see "Why are the Digital Humanities So White? or Thinking the Histories of Race and Computation" in the same volume <http://dhdebates.gc.cuny.edu/debates/text/29>.

[3] It is worth noting that one of the very few publications to deal with the EEBO set of projects in this way was published in *Literary and Linguistic Computing* (now *Digital Scholarship in the Humanities*). See Diana Kichuk, "Metamorphosis: Remediation in *Early English Books Online (EEBO)*" (2007) 22 (3): 291-303. DOI: http://dx.doi.org/10.1093/llc/fqm018. Kichuk's efforts have helped establish a historical framework for this discussion; this presentation seeks to contextualise the facts she has brought together and extend their relevance into discourses about DH as a whole.

Projet Odysseus, Outil d'Etudes Comparatives Du Traductologue

Marianne Reboul
odysseuspolymetis2010@gmail.com
Labex OBVIL, Université Paris-IV Sorbonne, France

Yuri Bizzoni
yuri.bizzoni@gmail.com
CLASP, Gothenburg University, Sweden

Introduction

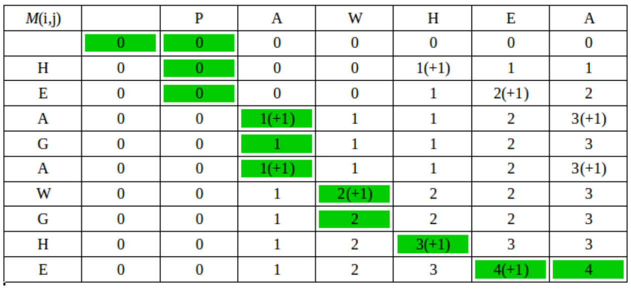
Le programme Projet Odysseus permet de voir immédiatement des différences et similitudes entre diverses traductions de l' Odyssée d'Homère. Nous appliquons ce programme à un corpus français d'une centaine de traductions françaises allant de 1584 à 1934, afin de déceler les évolutions diachroniques dans les pratiques traductives. Nous constituons un corpus italien similaire, et avons aussi effectué des essais sur des traductions espagnoles, anglaises et allemandes de l' Odyssée. Le protocole est applicable à tout corpus de traductions (Callison-Burch, 2008).

Notre programme permet donc de comparer un nombre indéfini de traductions d'un même texte source, sans dépendre des particularités linguistiques de chacune des langues cible.

Preprocessing: l'alignement

Pour comparer les différentes traductions entre elles, nous déterminons leurs similarités sémantiques et associons le texte source avec chaque texte cible. Nous établissons des points d'ancrage en grec et en français, qui permettent de découper les textes en séquences ayant potentiellement le même sens (Feng et Manmath, 2006) : les noms propres. Ils présentent deux avantages: ils peuvent être de très basse fréquence et sont majoritairement traduits par l'ensemble des traducteurs. Le texte grec pivot comprend toutes les leçons éditoriales établies à ce jour, afin d'assurer un alignement optimal tant avec les traductions anciennes que modernes. Le texte source pivot est découpé en séquences fixes : le programme découpe une nouvelle séquence dont le premier mot est le nom propre. Nous obtenons une série fixe de séquences. Nous appliquons le même processus aux textes cibles. Sont ainsi déterminés un ensemble de séquences sources et un ensemble de séquences cibles de nombre inégal.

Le programme aligne les séquences cibles aux séquences sources grâce à l'algorithme de Needleman-Wunsch (Needleman et Wunsch, 19 70). Il remplit une matrice de scores de similarité en comparant chaque séquence source avec chaque séquence cible. Un dictionnaire de noms propres augmente ou diminue le score. Un transformateur phonétique, transcrivant le grec en alphabet latin, augmente le score si le nom source phonétique se trouve dans la séquence cible. Le score tient compte la différence de longueur entre la source et la cible. Nous établissons dictionnaire de fréquence : si le nom pivot est à fréquence équivalente entre la source et la cible, le score augmente. Lorsque la matrice de scores est remplie, l'algorithme établit le chemin le plus probable depuis les deux dernières séquences source et cible jusqu'aux deux premières (Fig.1).



M(i,j)		P	A	W	H	E	A
	0	0	0	0	0	0	0
H	0	0	0	0	1(+1)	1	1
E	0	0	0	0	1	2(+1)	2
A	0	0	1(+1)	1	1	2	3(+1)
G	0	0	1	1	1	2	3
A	0	0	1(+1)	1	1	2	3(+1)
W	0	0	1	2(+1)	2	2	3
G	0	0	1	2	2	2	3
H	0	0	1	2	3(+1)	3	3
E	0	0	1	2	3	4(+1)	4

Fig. 1: Parcours de l'algorithmeTableauNW.png

Nous constituons ensuite un dictionnaire basé sur la distribution sémantique des mots. Chaque mot est caractérisé par un vecteur de cooccurrence : sur l'ensemble des séquences, lorsque un mot apparaît dans une fenêtre constituée d'une séquence source et de sa séquence cible, la valeur du vecteur à l'indice de la fenêtre augmente. Statistiquement, des mots à usage similaire se trouvent à la fois dans les séquences sources et cibles aux mêmes indices. La similarité des vecteurs est établie avec la similarité cosinus (Ye, 2011). Si un mot source obtient un vecteur similaire à celui d'un mot cible, la source-clé et la cible-valeur sont ajoutés au dictionnaire. Un second alignement est produit, tenant compte des scores obtenus grâce au dictionnaire de distribution. Lorsque nous effectuons plusieurs alignements en même temps, nous fusionnons les dictionnaires. Chaque séquence source comporte alors un identifiant unique auquel un identifiant cible est associé.

Classes comparatives

Nous pouvons comparer chacune des traductions alignées au texte source pivot, dans une interface faisant figurer autant de traductions que l'utilisateur le souhaite.

Lorsque des vides sont laissés par l'algorithme, deux interprétations sont possibles : le traducteur est extrêmement loin du texte source, ou il ne traduit pas le texte source. Dans les deux cas, cela nous renseigne sur ses pratiques traductives.

Nous comparons ensuite statistiquement les séquences alignées. Pour chaque séquence dans l'ensemble du corpus, nous déterminons la fréquence de chaque mot. Nous pouvons voir le taux d'usage d'un mot pour chaque séquence. Il est ainsi possible de visualiser les hapax produits par chaque traducteur. Il est aussi possible de déceler si un traducteur se base sur la même source que les autres : si le nombre d'hapax sur une séquence se multiplie, le traducteur ajoute probablement une séquence par rapport à ses homologues. Nous identifions les étoffements et réductions : si le nombre de mots employé par un traducteur pour une séquence est largement inférieur ou supérieur au nombre médian de mots employés dans le corpus pour cette séquence, le traducteur étoffe ou réduit le texte source. Nous décelons enfin un littéralisme fréquentiel de chaque séquence par rapport à sa source (lorsque le nombre de mots grecs est équivalent au nombre de mots français pour une même séquence), ainsi qu'un littéralisme phonétique (lemme grec retranscrit phonétiquement proche d'un mot cible).

Nous nous appuyons sur les travaux d'Henri Meschonnic (Meschonnic, 1997) et d'Antoine Berman (Berman, 1984) pour définir des critères de comparaison entre la source et les cibles. Nous déterminons d'abord s'il y a, entre les deux, un allongement ou un appauvrissement quantitatif, en définissant une fenêtre de mots au-delà ou en deçà de laquelle nous considérons que la différence dans le nombre de mots est signifiante. Nous visualisons la tendance à la « rationalisation » : nous repérons, en comparant les patrons syntaxiques, l'inversion de la linéarisation des arborescences syntaxiques. Nous visualisons enfin la « destruction des rythmes », en recherchant la similarité entre les patrons de ponctuation ou les patrons de syntaxe profonde (juxtaposition, coordination, subordination) entre la source et la cible.

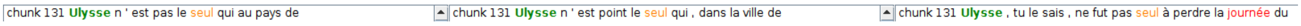
Pour comparer les traductions entre elles, l'utilisateur peut considérer dans l'étude statistique soit les traductions à l'écran, soit l'ensemble du corpus.

Nous intégrons un graphique permettant de visualiser les tendances décrites des auteurs dans la diachronie. Si un auteur a une forte tendance à l'emploi d'hapax, et que cette tendance se poursuit chez ses successeurs, cela est immédiatement visible.

Cas d'étude

Voici deux exemples: une utilisation unilingue dans l'interface de comparaison automatique, et une utilisation multilingue avec les résultats d'un dictionnaire de distribution. Le premier exemple est basé sur le chant I de vingt traductions françaises, de 1584 à 1935.

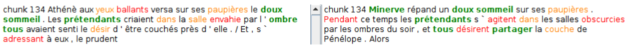
Fig. 2: Une des fenêtres du programmescreen2.png



chunk 131 Ulysse n ' est pas le seul qui au pays de | chunk 131 Ulysse n ' est point le seul qui , dans la ville de | chunk 131 Ulysse , tu le sais , ne fut pas seul à perdre la journée du

Fig. 3: Trois séquenceszoom1.png

Dans la Fig.3, la première version est de Médéric Dufour et Jeanne Raison (Dufour et Raison, 1935), la deuxième celle d'Eugène Bareste (Bareste, 1842), la troisième celle de Victor Bérard (Bérard, >1924). Les termes en vert sont utilisés pour chaque tronçon par une grande majorité de traducteurs, ceux en orange sont fréquents, et ceux en rouge sont des hapax (utilisés une seule fois dans cette séquence par un traducteur). Le programme permet de visualiser ces captures sur une grande échelle, avec autant de traducteurs que nécessaire. En comparant le « chunk 131 » de Victor Bérard, par exemple, avec l'ensemble des « chunk 131 » du corpus, nous distinguons, toute proportion gardée, une tendance de Bérard à se démarquer de ses prédécesseurs par emploi de termes rares, phonétiquement ou syntaxiquement proches du grec.

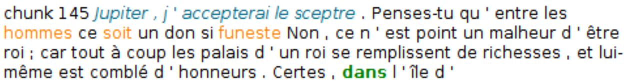


chunk 134 Athénè aux yeux ballants versa sur ses paupières le doux sommeil . Les prétendants criaient dans la salle envahie par l ' ombre tous avaient senti le désir d ' être couchés près d ' elle . / Et , s ` adressant à eux , le prudent

chunk 134 Minerve répand un doux sommeil sur ses paupières . Pendant ce temps les prétendants s ` agitent dans les salles obscurcies par les ombres du soir , et tous désirent partager la couche de Pénélope . Alors

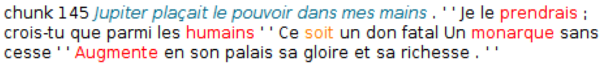
Fig. 4: Comparaison, zoom

Dans la Fig. 4 Dufour et Raison (premier texte) optent pour un terme rarement associé aux yeux d'Athéna (« ballants »), produisant un hypallage, tandis que les autres termes sont consensuels. Bareste privilégie les présents d'hypotypose et de juxtaposition temporelle. Certaines expressions homériques sont rendues unanimement par les traducteurs, comme « ὕπνος ἡδύς », « le doux sommeil ».



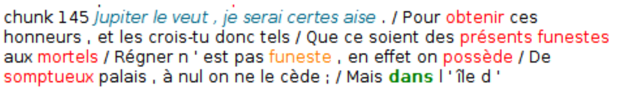
chunk 145 *Jupiter , j ' accepterai le sceptre* . Penses-tu qu ' entre les
hommes ce soit un don si funeste Non , ce n ' est point un malheur d ' être
roi ; car tout à coup les palais d ' un roi se remplissent de richesses , et lui-
même est comblé d ' honneurs . Certes , dans l ' île d '

Fig. 5: Traduction de Bareste (1842)



chunk 145 *Jupiter plaçait le pouvoir dans mes mains* . ' ' Je le prendrais ;
crois-tu que parmi les humains ' ' Ce soit un don fatal Un monarque sans
cesse ' ' Augmente en son palais sa gloire et sa richesse . ' '

Fig. 6: Traduction de Bignan (1853)

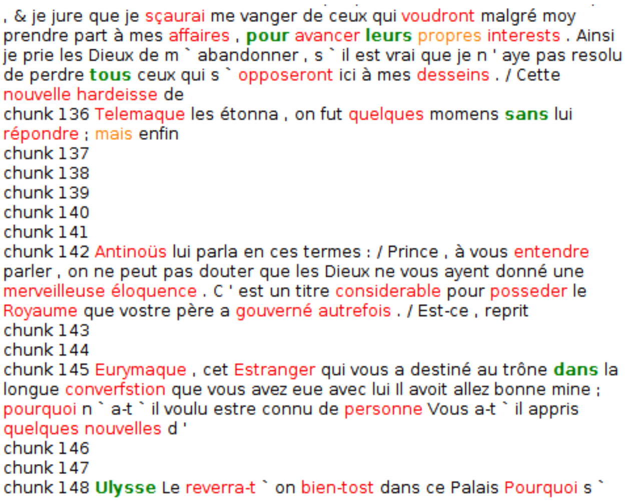


chunk 145 *Jupiter le veut , je serai certes aise* . / Pour obtenir ces
honneurs , et les crois-tu donc tels / Que ce soient des présents funestes
aux mortels / Régner n ' est pas funeste , en effet on possède / De
somptueux palais , à nul on ne le cède ; / Mais dans l ' île d '

Fig. 7: Traduction de Froment (1883)

Les Fig. 5, Fig. 6 et Fig. 7 sont des traductions d'une portion de texte qui tend à se traduire selon le même mode au XIXe siècle. Dans plus de vingt textes, pour l'expression « καὶ κεν τοῦτ' ἐθέλοιμι Διός γε διδόντος ἀρέσθαι », les traducteurs imitent la ponctuation du texte source (italique), mais aussi sa syntaxe (couleur bleue).

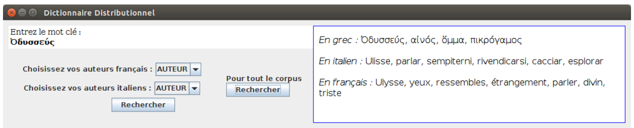
Le programme permet de déceler les portions de texte omises par le traducteur. Par exemple, Achille de La Valterie disait traduire l' Odyssée sans connaître le grec (à partir du latin) : sa traduction s'aligne, mais le nombre d'hapax et d'omissions est sans précédent (Fig. 8).



, & je jure que je sçaurai me vanger de ceux qui voudront malgré moy
prendre part à mes affaires , pour avancer leurs propres interests . Ainsi
je prie les Dieux de m ` abandonner , s ` il est vrai que je n ' aye pas resolu
de perdre tous ceux qui s ` opposeront ici à mes desseins . / Cette
nouvelle hardeisse de
chunk 136 Telemaque les étonna , on fut quelques momens sans lui
répondre ; mais enfin
chunk 137
chunk 138
chunk 139
chunk 140
chunk 141
chunk 142 Antinoüs lui parla en ces termes : / Prince , à vous entendre
parler , on ne peut pas douter que les Dieux ne vous ayent donné une
merveilleuse éloquence . C ' est un titre considerable pour posseder le
Royaume que vostre père a gouverné autrefois . / Est-ce , reprit
chunk 143
chunk 144
chunk 145 Eurymaque , cet Estranger qui vous a destiné au trône dans la
longue converfstion que vous avez eue avec lui Il avoit allez bonne mine ;
pourquoi n ` a-t ` il voulu estre connu de personne Vous a-t ` il appris
quelques nouvelles d '
chunk 146
chunk 147
chunk 148 Ulysse Le reverra-t ` on bien-tost dans ce Palais Pourquoi s `

Fig. 8: Traduction de La ValterieValterie

Dans cet extrait, les blancs laissés par le programme ne sont pas des erreurs d'alignement, mais des omissions du traducteur, qui choisit de ne pas traduire.


Dictionnaire Distributionnel
Choisissez vos auteurs français : AUTEUR
Choisissez vos auteurs italiens : AUTEUR
Rechercher
Pour tout le corpus
Rechercher

Entrez le mot clé :
Ὀδυσσεύς

En grec : Ὀδυσσεύς, αἰνός, ὄμμα, πικρόγαμος

En italien : Ulisse, parlar, sempiterni, rivendicarsi, cacciar, esplorar

En français : Ulysse, yeux, ressembles, étrangement, parler, divin, triste

Fig. 9: Dictionnaire

Pour une utilisation multilingue, nous pouvons mener les mêmes expériences que pour un corpus unilingue. Nous pouvons aussi visualiser le dictionnaire de distribution, qui permet de voir comment chaque langue traduit, pour l'ensemble de son corpus, chaque notion. En Fig. 9, nous comparons comment chaque langue traduit « Ὀδυσσεύς » dans le chant I et quels en sont les termes sémantiquement proches, tous auteurs confondus. Nous pouvons ne choisir que certains auteurs, afin par exemple de voir l'évolution de l'emploi d'un terme en France et en Italie pour une période définie. Nous pouvons générer des dictionnaires multilingues dont le degré de sévérité dans la sélection est modifiable.

Conclusion

Ce programme a pour but de proposer aux comparatistes une façon simple de repérer les grandes tendances et les phénomènes notables d'une traduction à une autre, indépendamment de la langue du traducteur. Mais le programme ouvre aussi la voie à la création d'autres outils pour l'étude de corpus multilingues (dictionnaires, détection automatique de paraphrases interlinguistiques), à de nombreuses analyses statistiques (fréquence des mots, « type / token ratio », etc.) qui représenteront un apport substantiel à l'étude de l'Histoire des traductions.

Bibliography

Bareste, E. (1842). *Odyssée: traduction nouvelle*, (Lavigne).

Bérard, V. (1924). *L'Odyssée. Texte établi et trad. par Victor Bérard*, (Les Belles Lettres).

Berman, A. (1984). *L'épreuve de l'étranger: Culture et traduction dans l'Allemagne romantique* Vol. **226**: 67-81, Gallimard.

Berman, A. (1995) *Pour une critique des traductions: John Donne*. Paris, Éditions Gallimard, "Bibliothèque des idées".

Bignan, A. (1853). *L'Odyssée traduite en vers français*, (Hachette).

Callison-Burch, C. (2008). *Paraphrasing and translation* (Doctoral dissertation, University of Edinburgh).

Dufour, M., Raison, J. (1941). Homère. *L'Odyssée: traduction de Médéric Dufour et Jeanne Raison, illustrée de vingt-quatre compositions en couleurs par Benito...* ,(Le Vasseur).

Feng, S. and Manmatha, R. (2006, June). A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on* , IEEE, pp. 109-18

Froment, J. B. F. (1883). *Odyssée d'Homère*, Vol. **2**: in-8e (Plon et Cie)

La Valterie, A. (1681). *L' Odyssée*, (Barbin)

Meschonnic, H. (1999). *Poétique du traduire*. Paris: Verdier.

Ye, J. (2011). Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Mathematical and Computer Modelling*, **53**(1), 91-97.

The Lifecycle of a Digital African Studies Projects: Creating Sustainable, Equitable, and Ethical Projects

Dean Rehberger
dean.rehberger@matrix.msu.edu
Michigan State University, United States of America

Ibrahima Thiaw
thiawi@yahoo.com
Institut Fondamental d'Afrique Noire (IFAN), University Cheikh Anta Diop of Dakar, Senegal

Deborah Mack
MackDLynn@si.edu
Smithsonian National Museum of African American History and Culture

Candace Keller
kellercm@cal.msu.edu
Michigan State University, United States of America

Catherine Foley
Catherine.Foley@matrix.msu.edu
Michigan State University, United States of America

For more than 20 years, researchers at Matrix, the Center for Digital Humanites and Social Science have been working on digital projects in several countries in Africa. While the technologies are critical parts of the digital humanities, ethical considerations also need to be part of any project that involves multiple projects. This is particularly true of Digital African Studies Projects because of the long and bloody history of colonialism, exploitation, and cultural theft. This long paper will explore through the context of two ongoing projects -- "Archive of Malian Photography" and the "Gorée Island Archaeological Digital Repository" — strategies to be deployed to develop sustainable, equitable, and ethical projects. While neither project is a perfect model, the strategies deployed set against the everyday frustrations of multiple partner projects, long distance project management, and problematic working conditions does help to expose what works and what still needs to be changed or augmented.

Mali has remained the international nexus of African photography for nearly twenty years. Since 1994, its capital, Bamako, has been home to the only biennial festival of photography from Africa and has produced the continent's most globally renowned professional photographers. Matrix in collaboration with the Maison Africaine de la Photographie in Mali, is digitizing and rendering globally accessible the archives of the nation's most important photographers, dating from the 1940s to the present. The

"Archive of Malian Photography" addresses the following significant needs for scholarship and teaching in the humanities as well as the preservation of Mali's cultural heritage:

- Although commercial, popular, and scholarly interest in African photography has flourished over the past twenty years, access to photographers' studio archives is extremely limited and published materials are minimal;

- Retained in private archives, these materials are not catalogued, appropriately preserved, or internationally accessible for research and education;

- Due to the high commercial value of these archives in global markets, these materials are vulnerable to mistreatment, theft, and exploitation;

- Moreover, stored in harsh climactic conditions, the physical integrity of these collections is in serious jeopardy.

Photography was introduced to present-day Mali during the 1880s by French military officers and, later, colonial administrators, missionaries, and French expatriates. By the 1940s, an African market for photography developed in the French Sudan, as Mali was then known, and its professional photographers maintained a monopoly over the medium. As a result, their archives contain rare visual documentation of social, cultural, and political life as well as processes of urban development in the country, and in French West Africa more broadly. Spanning the eras of French colonialism, political independence, socialism, and democracy, their archives record important sociopolitical transformations of present-day Mali, its capital, and smaller towns along the Niger River such as Mopti and Ségu during the twentieth century. Employed by colonial and national governments, while operating private studio enterprises, each photographer's collection houses unique African perspectives on local histories and practices, including personal and family portraiture, military activities, visits of foreign dignitaries, and images of the coup d'états that toppled the socialist regime of the nation's first president, Modibo Keïta (1968), and the dictatorial rule of his successor, Moussa Traoré (1991). They also feature the construction of national monuments, governmental structures, bridges, dams, roadways, as well as prominent religious leaders, political figures, cultural ceremonies, and fluctuating trends in personal adornment, popular culture, and photographic practices from the 1940s to the present.

In addition to promoting the global accessibility of these materials, the project is designed to protect the physical integrity of the original archives, safeguarding the negatives from further damage and theft. After the popularization of Malian photographers' images in exhibitions and publications around the world since the 1990s, their archives have become increasingly vulnerable to pilfering and misuse due to their high commercial value in the international art market. Underscoring the severity of this issue, today, some images from the archives of El Hadj Tijani Sitou and Abdourahmane Sakaly are illegally featured for sale online (http://www.african-collection.dk/english/sakaly-1.htm) by a dealer in Denmark who obtained the photographers' original negatives, along with those from Mamadou Cissé's archives, from an administrator at the National Museum in Bamako who had been charged with their protection. To date, the dealer refuses to repatriate the negatives. In another case, prints from twelve of Sitou's negatives appeared on the cover of African Arts in 2008, without his family's permission. Such pervasive pilfering of negative archives by foreign collectors, dealers, curators, and scholars, as well as by the leaders of local cultural institutions (museums, galleries, libraries), has prevented these rare collections from entering formal, public archives in Mali.

The Gorée Island Archaeological Digital Repository seeks is creating virtual 3D representations of cultural heritage materials excavated from archaeological sites in and around Gorée Island, Senegal and share those representations in an open-access, online digital repository. This work answers at least three needs in the scholarly community. First, it enables cultural heritage institutions in Africa to digital preserve archaeological materials in a way that is unobtrusive, relatively inexpensive. The project makes use of photogrammetry, a process wherein a series of images are taken of an object and run through point-recognition software to create a virtual 3D image. This process inexpensive, portable, and relatively easy to complete, making it the jointly sustainable choice for African cultural heritage institutions. Second, the repository allows scholars both within Africa and around the world to have access to these cultural heritage materials without the restrictions and inconveniences of analog-based preservation. Finally, the Gorée Island Digital Repository builds capacity amongst Senegalese students, scholars, and organizations to continue documenting and preserving cultural heritage materials using industry best standards and practices.

The project is being developed through the collaboration of the Université Cheikh Anta Diop de Dakar – Sénégal (UCAD), Institut Fondamental d'Afrique Noire, the Smithsonian National Museum of African American History and Culture, Michigan State University, Matrix with the support of the Centre de Researche Ouest Africain.

Documenting, safeguarding, preserving, interpreting/reinterpreting, and making accessible the myriad expressions of Africa's many cultures is vitally important for Africa's diverse constituent communities as well as for the rest of the world. Museums, libraries, and archives in Africa and around the world face an enormous challenge as Africa's diverse and rich cultural heritage has been scattered by history and put at risk by wars, illicit trafficking, overwhelming economic challenges, and destruction or erosion due to human and environmental impacts. As pioneering efforts have demonstrated, the digital revolution opens up significant possibilities for long-term preservation and meaningful access that have proven unattainable in an analog world. However, to do so requires thoughtful, ethical, equitable, and sustainable strategies. These considerations need to be built into the very structure of digital humanities projects.

Significant barriers remain, however, that severely limit efforts to move beyond these individual projects and fully utilize digital technologies within Africa and with African cultural materials around the world. While we are at a moment of opportunity with digital technologies, this is also a period of crisis where invaluable cultural resources are at risk to be lost forever. This paper also calls for the need for major international leadership initiatives to tackle these barriers in coordinated fashion and chart a path forward for museums, libraries, archives, universities and other heritage preservation institutions to construct equitable international partnerships that can harness this powerful opportunity.

Addressing Torture in Iraq through Critical Digital Media Art—Hearts and Minds: The Interrogations Project

Scott Rettberg
scott.rettberg@uib.no
University of Bergen, Norway

Roderick Coover
roderickcoover@gmail.com
Temple University

Hearts and Minds: the Interrogations Project is an interactive virtual reality artwork developed by an interdisciplinary team including humanists, social scientists, artists, and computer scientists from four different universities. The project attempts to extend and make accessible difficult narratives of war and abusive violence based on actual accounts from soldiers involved. The work offers models for engaging with testimony and oral history. It uses visualization to build new discourse around challenging topics and to bridge concepts that enable storytelling. While many uses of visualization technologies are focused on providing accessible representation of "big data," in this case, the technologies are being used as a narrative platform to represent a complex contemporary issue and to provide a platform for discussion and debate of military interrogation methods and their effects on detainees, soldiers, and society.

Hearts and Minds makes use of the CAVE2^{TM} environment for a multisensory artwork addressing a complex contemporary problem: as American soldiers are returning from wars in Iraq and Afghanistan, it is becoming increasingly clear that some of them participated in interrogation practices and acts of abusive violence with detainees for which they were not properly trained or psychologically prepared. This has in turn left many soldiers dealing with Post-Traumatic Stress Disorder after their return home. More than 1.8 million US troops have served in Iraq and Afghanistan, with 37% having been deployed at least twice (Litz, 2009) The mental health impact of these wars is still under research as many veterans are at risk for chronic PTSD. American soldiers and citizens are left with many unresolved questions about the moral calculus of using torture as an interrogation strategy in American military operations.

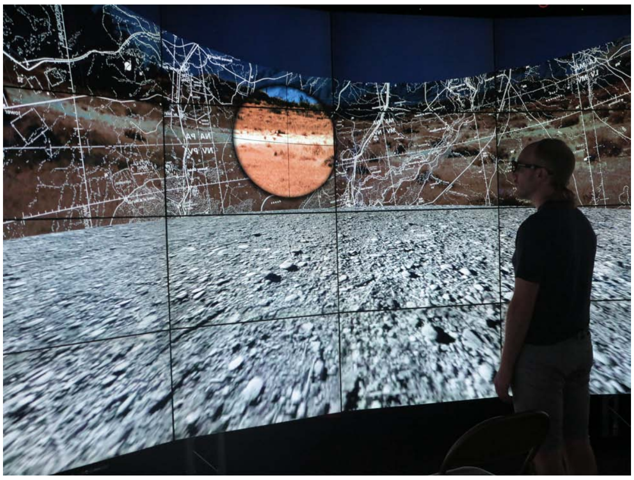


Fig 1. A performer interacts with the CAVE2 version of *Hearts and Minds* at the EVL in Chicago

Development

Hearts and Minds bridges art, computer science, and social science research. Artist Roderick Coover (Temple University) and writer Scott Rettberg (University of Bergen) worked with the research scholars John Tsukayama and Jeffrey Stevenson Murer (St. Andrews University) to distill central themes and stories from the significant and extensive research project—based on hundreds of hours of original interviews with veterans—carried out by Tsukayama (Tsukayama, 2014). These interviews include revelations of a highly sensitive nature, including narratives of participation in acts of abusive violence that entailed violations of human rights. The interviewees granted Tsukayama the right to use their stories in his dissertation and in subsequent research outcomes derived from it, provided that their identities remained anonymous. The tapes of recorded interviews were destroyed after transcription, except for short samples to prove their authenticity, and Tsukayama did not retain any personal contact information for the soldiers he interviewed. The text was condensed into an accessible and coherent set of stories that would preserve the accuracy of the testimonies while voice actors would perform the roles of veterans, further assuring their anonymity.

Coover and Rettberg worked with artist Daria Tsoupikova and scientist Arthur Nishimoto at the Electronic Visualization Lab (University of Illinois at Chicag, 2014) to transform this controversial and challenging research into an accessible form through visualization and dramatization. Together they developed an interactive virtual environment with imagery, 3D models and panoramic photographic backgrounds to bring story elements together. Working across these environments allowed new kinds of connections to be made between home spaces and battlefields, and between domestic objects and the memories they become attached to.

The voice recordings performed by Philadelphia-based actors were integrated with interactive media elements in a peformative interactive environment. Objects, 3D environments, textures, and some animations were developed in Maya (Autodesk Inc., CA). Maya speeds up the production process through its rich selection of tools supporting all stages of modeling, including surface creation and manipulation, texturing, lighting, rigging, and animation. The visual, auditory and narrative elements were brought together in the Unity platform (Unity Technologies Inc., CA), software that is typically used by computer game developers. The getReal3D plugin for Unity developed by Mechdyne Corporation is used to run Unity across the CAVE2™ cluster (Mechdyne Corporation, 2014). User interaction was scripted using the Omicron (Electronic Visualization Lab, 2014) input abstraction library developed by EVL.

CAVE2 Performance

In its first iteration, *Hearts and Minds* was presented as public performances at the University of Illinois Chicago Electronic Visualization Lab in June and July 2014 (Galatzer-Levy, 2014). Chicago-based performance artist Mark Jeffrey led a performance of the interactive work. As the audience entered the space, they found themselves in a temple environment listening to each of the four soldier character's stories of enlistment—why they originally chose to become soldiers and what motivated their perspectives on the purpose for military service.

Jeffery then led the audience to the boy's room, where activating four individual objects launches stories of first encounters with abusive violence in military experience, such as in hazing rituals during basic training, or after first arriving in Iraq. When each trigger object is selected within the 3D visual space, the domestic space falls away and a surreal desert landscape is revealed. This transition serves as a metaphor for the interior state of the individual solider, as it is coherent with accounts of soldiers experiencing Post-Traumatic Stress. It is also intended to bring audience members into a "listening state" where they can focus on the individual voices and the issues they raise. Objects in a living room space and a suburban backyard move us further into the field of battle, and there we encounter harrowing stories of interrogation, torture, and moral conflicts confronted differently by each of the characters.

An important component of the performances of *Hearts and Minds* is that the experience of the artwork is followed by audience discussion. The ultimate purpose of this work is to promote dialogue and debate about the contexts and circumstances of the use of torture in battlefield torture in recent history. In this sense the project shares an aim with Augosto Boal's Theatre of the Oppressed, in that attempts to offer audiences "the aesthetic means to analyze their past, in the context of their present, and subsequently to invent their future, without waiting for it." (International Theatre of the Oppressed Organization, 2014) During a 2015 presentation of the project at the Oslo Human Rights Film Festival (Coover et al., 2015), discussion participants included both a veteran police interrogator, who shared his experience that more humane methods of interrogation than torture were universally more effective, and a number of prisoners of conscience who had themselves been subjec to abusive violence and felt provoked by the work to share their own experiences as victims of torture.

Accessible Versions and Alternative Platforms

One of the challenge for makers of immersive virtual reality artworks, particularly those developed in CAVEs and other custom-built VR environments is that these artworks tend to be more often read about in the literature of the field than experienced first-hand. The CAVE2™ at the EVL for example is an active research lab facility with keycard access at the center of a large engineering building on the UIC campus. *Hearts and Minds* has been shown there at several special events and on specially arranged tours, but it is not the ideal accessiblity situation to reach a broader audience. Because *Hearts and Minds* was developed in Unity, it is possible to port the application to other environments. In order to make the work more accessible for new audiences, the *Hearts and Minds* team is developing new versions of the work: a Mac OS standalone application suitable for cinematic performance or installation (Coover et al., 2014), a Unity web-player version which will be published on the World Wide Web and will be accessible freely to the public, a version suitable for iPads and other tablet computers, and a version for Oculus Rift. The stand-alone application version has been used for performances in cinema enivronments in Paris, Bergen, Oslo, Krakow, Michigan, and elsewhere, and the other application versions will be publically released in 2016.

Fig. 2 Presentation of the interactive cinematic version of *Hearts and Minds* at the 2015 Oslo Human Rights / Human Wrongs festival

Completing *Hearts and Minds* as Critical Digital Media and Digital Humanities Project

The *Hearts and Minds* team is currently developing the tablet version of the work, which will include both the tablet (iPad and Android) application, and a package of contextualizing essays, interviews and commentaries both specific to the work and that provide more detailed treatments of the issues the work raises. These essays and interviews will for example contributions by political scientists and social scientists about the situations out of which abusive interrogations practice emerged, contributions by digital humanists considering interactive media artworks as a mode of critically and socially engaged research, and contributions by designers and computer scientists considering the history and technologies involved in virtual reality art. Coover and Rettberg's long paper for DH 2016 will include a presentation of this new version of the project, and will further situate *Hearts and Minds* in an emergent corpus of critically engaged new media art.

Bibliography

Coover, R., Nishimoto, A., Rettberg, S. and Tsoupikova, D. (2014). Hearts and Minds: The Interrogations Project. *Proceedings of the IEEE VIS Arts Program* (VISAP), Paris, France.

Coover, R., Murer, J., Kvanvig, G., Rachlew, A., Rettberg, S. and Tsukayama, J. (2015). Interrogating Torture. Video of panel discussion at 2015 Oslo Human Rights Festival. https://vimeo.com/119752276 (accessed May 24, 2015).

Electronic Visualization Lab (2014). Omnicron. http://github.com/uic-evl/omicron (accessed May 24, 2015).

Galatzer-Levy, J. (2014). Virtual immersion in 3-D storytelling. *UIC News*. http://news.uic.edu/virtual-immersion-in-3-d-storytelling (accessed May 24, 2015).

International Theatre of the Oppressed Organization (2014). Declaration of Principles. http://www.theatreoftheoppressed.org/en/index.php?nodeID=23 (accessed March 5, 2016).

Litz, B. and Schlenger, W. (2009). PTSD in Service Members and New Veterans of the Iraq and Afghanistan Wars: A Bibliography and Critique, *PTSD Research Quarterly*, **20**(1): 1-7.

Mechdyne Coporation (2014). getReal3D. http://www.mechdyne.com/getreal3d.aspx (accessed May 24, 2015).

Tsukyama, J. (2014). *By Any Means Necessary: An Interpretive Phenomenological Analysis Study Of Post 9/11 American Abusive Violence In Iraq*. Ph.D. thesis, University of St Andrews.

An Iterative 3DGIS Analysis of the Role of Visibility in Ancient Landscapes

Heather Richards-Rissetto
richards-rissetto@unl.edu
University of Nebraska-Lincoln, United States of America

Introduction

Through the spatial arrangements of temples, houses, roads, and more, the built environment provides a window to human interaction (Lawrence and Low, 1990). Spatial configurations influence how people negotiate their surroundings. People "read environmental cues, make judgments…and then act accordingly" (Rapoport, 1990: 139), and these decisions in turn, affect the frequency and intensity of interaction (Fletcher 1981). While many factors influence interaction within landscapes, in this paper I focus on visibility.

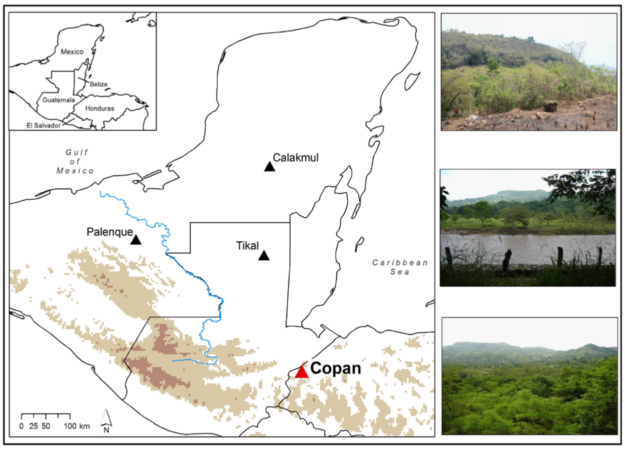


Figure 1: Map of Copan's location at southeastern periphery of Maya region (Map: H. Richards-Rissetto)

The visibility, intervisibility, and invisibility of features communicate information that guides pedestrian movement, and consequently, structures social interaction and community organization (Llobera, 2003, 2006; Gillings, 2015). Building on these ideas, this paper uses Geographic Information Systems (GIS) and 3D visualization to explore the role of visibility in ancient landscapes asking: *How might visibility influence where people went, what they did, who interacted with whom, and how did these interactions shape their daily experiences?*

Case Study: Copan, Honduras

The case study is the ancient Maya polity of Copan (Figure 1) ruled for over four-hundred years by a line of dynastic kings who by the late eighth century were facing mounting sociopolitical and environmental problems (Fash, 2001). Copan's final dynastic ruler, *Yax Pasaj*, like the rulers of many other Maya polities, was coping with strenuous environmental, demographic, and sociopolitical circumstances that would ultimately lead to the kingdom's demise. Yet during this time of stress, it seems that *Yax Pasaj* carried out a major urban renewal project commissioning several new temples in the city center that elevated Copan's skyline. Given the changes to Copan's urban fabric, *Yax Pasaj's* reign is an ideal case study to investigate the role visibility may have played in the production or reproduction of social interaction among the ancient Maya.

At Copan, as at other Maya centers, imagery on ceramics, murals, and freestanding monuments depicted deities floating over lords who subsequently looked down over lower-ranking persons. Maya architecture replicated this vertical succession by elevating royal compounds above other architecture, and in essence linking Maya rulers to the heavens (Messenger, 1987). In terms of visibility, epigraphic decipherments indicate that "seeing" afforded high status, and sight had an authorizing gaze and witnessing function—similar to Foucault's (1995) Panoptic gaze—where those who were all-seeing were all-knowing (Houston et al. 2006: 173). In order to be all-seeing or to give such an impression, however, Maya rulers needed to be seen, and so often located themselves in physically high and easily visible places or built tall temples that dominated the landscape.

While we know that Maya kings typically constructed highly visible temples, we actually know very little about the role visibility may have played in structuring social connections and daily interactions among social groups. To do this we need to broaden our view from civic-ceremonial precincts to encapsulate the broader landscape (Doyle et al., 2012; Richards-Rissetto 2010; Landau, 2015).

Background: GIS & 3D Visualization

Early visibility studies in the Maya region focused on astronomical alignments among structures, freestanding monuments, and the sky (Aveni and Hartung, 1986). Later, ethnographic studies showing that contemporary Maya often use sight lines to mark out spaces (Hanks, 1990) inspired researchers to investigate whether non-astronomical lines-of-sight also existed at ancient sites; and in fact, archaeologists identified sight lines between a major temple and outlying stelae at the site of La Milpa, Belize (Hammond and Tourtellot, 1999). Recent research has moved away from lines-of-sight between two objects to study the relationships that an object may have to the many objects or features found within a landscape, referred to as a visualscape (Llobera, 2003). Simple line-of-sight measurements cannot provide data on the relationships among multiple objects because they are done along a fixed line; however, visualscapes can be measured using viewsheds that calculate an object's entire 360° field-of-view using GIS.

A GIS links mapped features to attributes stored in a database and overlays different data layers such as land usage, elevation, and buildings to help reveal complex patterns, relationships, and trends that are not readily apparent using other tools such as traditional databases not linked to maps.

Pros: In regard to visibility analysis, GIS allows archaeologists to move beyond line-of-sight analysis to viewshed analysis. A viewshed uses raster data (pixels) to identify all cells visible from one or more viewpoints in a landscape; all non-visible cells are assigned a 0 and all visible cells are assigned a 1. This basic binary schema allows for complex mathematical calculations, for example, Boolean operations or map algebra, to calculate topographic prominence of individual features (or classes of features) and percentage of intervisibility among features.

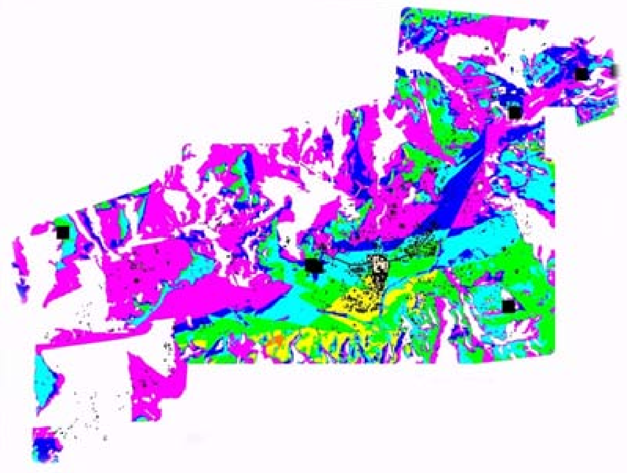


Figure 2: Cumulative viewshed illustrating number of valley stelae visible at locations at Copan

Cons: "Viewsheds depicted in a GIS map bear little resemblance to what people experience on the ground" (Conolly and Lake, 2006: 233). This limitation occurs because viewshed data are 2.5D. In other words, viewsheds store heights and elevation, but they are not actually 3D models (Figure 2). For digital humanists, these flat maps lack a sense of mass, scale, and aesthetics integral to human perception and experience, and the numerical outputs fail to differentiate visibility of a building's façade versus its sides or back—essential for close reading interpretation. Technically, data resolution (i.e., ratio of pixel size to earth's surface) can dramatically affect viewshed results—low spatial resolution often masking variation and too high a spatial resolution underestimating visibility (King et al., 2015).

3D technologies offer an alternative to GIS. 3D data acquisition (e.g., airborne LiDAR, terrestrial laser scanning, and photogrammetry), 3D modeling (e.g., SketchUp, 3D StudioMax, Agisoft), and interactive 3D visualization (e.g., Unity, Oculus Rift) are transforming archaeological practice. But, what impact are such 3D technologies having on visibility analysis across ancient landscapes? Airborne LiDAR, for example, rapidly collects 3D data for archaeological sites across vast areas (Thompson and Prufer, 2015). Most LiDAR data are of unexcavated mounds requiring subsequent 3D modeling of architecture and proper alignment within terrains in order to perform visibility analysis—traditionally time-consuming tasks (Richards-Rissetto, 2013).

While most visibility analyses of archaeological landscapes use traditional 2.5D GIS, recently archaeologists have been exploring the potential of 3D approaches for visibility analysis in archaeology. Paliou (2014) developed a computational visibility approach to analyze the visual range of paintings first using 3D modeling programs (3DStudioMax and AutoCAD) and then converting the results into raster maps to be analyzed in a GIS. Dell' Unto and colleagues (2015) bring georeferenced 3D architectural models (using laser scanning and photogrammetry) into a GIS to calculate visibility of building interiors at Pompeii. While Saldana and Johanson (2015) also use 3DGIS, they employ procedural modeling to rapidly generate alternative 3D building reconstructions based on a set of architectural rules and attributes stored in a GIS to explore visibility in Ancient Rome (Saldana, 2015).

Methods

Building on this scholarship, I employ an iterative 3DGIS approach to explore the role of visibility at the ancient Maya site of Copan—today a UNESCO World Heritage Site in Honduras. The approach is twofold: computational and experiential. In the computational approach, I employ traditional 2.5D viewshed analysis in GIS to establish a baseline for comparative analysis with viewshed results in 3DGIS.

First, I use ArcGIS 10.3 (standard GIS software) to assign known building heights and interpolate building heights of unexcavated mounds and run viewsheds to calculate topographic prominence and percent visibility in relation to settlement of major temples and classes of architecture (Richards-Rissetto, 2013). Recent acquisition of airborne LiDAR data has generated a 1m resolution terrain allowing for greater accuracy than earlier visibility analyses (Richards-Rissetto, 2010; von Schwerin et al., 2016). Second, I employ CityEngine—a procedural modeling program that convert GIS data to 3D models—to generate 3D models for Copan's 3,000+ buildings with the LiDAR terrain using the GIS data and a set of architectural rules as well as laser scanned and photogrammetric models of some standing monuments at Copan (Figure 3) (Muller et al., 2006; Richards-Rissetto and Plessing, 2015; von Schwerin et al., 2013). These procedurally-generated 3D models are then returned to ArcScene (a 3D viewer for ArcGIS) and the viewshed analysis is rerun for comparative analysis of 2.5DGIS vs. 3DGIS of visibility at ancient Copan.

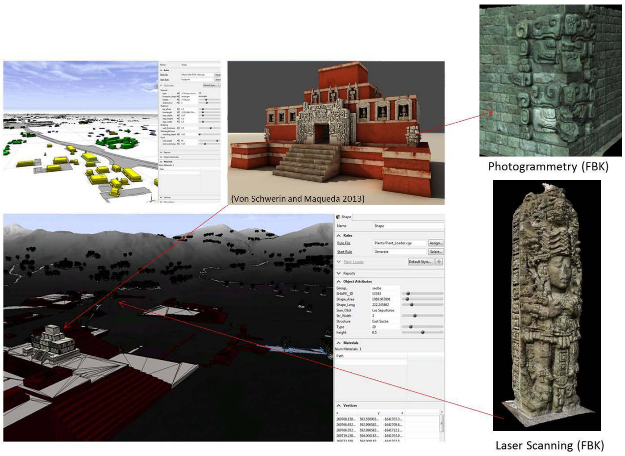


Figure 3: Illustrating procedurally-generated models and various data types imported into CityEngine

In the experiential approach, I export the 3D models and terrain from CityEngine into Unity 5—a gaming engine—to interactively explore the 3D models. In this model, vegetation is added to the landscape and avatars proceed along set paths generated from a combined cost surface and visibility analysis (Figure 4) (Richards-Rissetto, 2013; Richards-Rissetto and Landau, 2014). Oculus Rift—a head-mounted virtual reality display—is employed to create an immersive experience for ancient Copan as a means to more intuitively interact with archaeological data (Bartolo et al., 2000; Frisher and Dakouri-Hild, 2008).

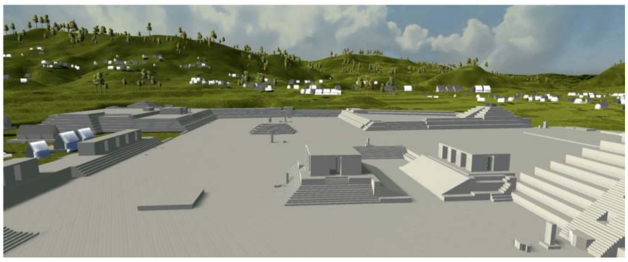


Figure 4: 3D Models (from SketchUp using GIS data) visualization in Unity 5 (Richards-Rissetto and Day)

Discussion

Strongly embedded in the Digital Humanities, this 3DGIS iterative approach tacks back and forth between

2.5D and 3D data to compare results and potentially derive new methods and interpretations for visibility analysis of ancient landscapes—analyses that would not be possible without taking advantage of the digital to cross-cut the computational and experiential.

Acknowledgements

The Layman Award, University of Nebraska-Lincoln provided a seed grant to carry out initial procedural modeling tests. This research would not be possible without permission and assistance from the Honduran Institute of Anthropology and History (IHAH). The MayaArch3D Project has generously providing the airborne LiDAR data and the laser scanned and photogrammetric models for this research. I want to thank UNL students Zachary Day, Stephanie Sterling, and Rachel Plessing for their important work on the visualizations.

Bibliography

Aveni, A., and Hartung, H. (1986). Maya City Planning and the Calendar. *Transactions of the American Philosophical Society*, 0065-9746, Vol. **76**, pt. 7. Philadelphia: American Philosophical Society.

Barcelo, J., Forte, M. and Sanders, D. (Eds.) (2000). *Virtual reality in archaeology.* Oxford: Archaeopress.

Conolly, J. and Lake M. (2006). *Geographical Information Systems in Archaeology*. Cambridge University Press.

Dell' Unto, N., Landeschi, G., Leander, T., Touati, A., Dellepiane, M., Callieri M. and Ferdani, D. (2015). Experiencing Ancient Buildings from a 3D GIS Perspective: a Case Drawn from the Swedish Pompeii Project. *Journal of archaeological method and theory*.

Doyle, J., Garrison, T. and Houston, S. (2012). Watchful Realms: integrating GIS analysis and political history in the southern Maya lowlands. *Antiquity*, **86**(333): 972-807.

Fash, W. (2001). *Scribes, Warriors, and Kings: The City of Copan and the Ancient Maya*. Thames and Hudson, London.

Fletcher, R. (1981). People and Space: A Case Study on Material Behavior. In *Pattern of the Past: Studies in Honour of David Clarke*, edited by I. Hodder, G. Issac, and N. Hammond, Cambridge University Press, Cambridge, pp. 97-128.

Foucault, M. (1995). *Discipline and Punishment.* Vintage Books, New York.

Frisher, B. and Dakouri-Hild, A. (Eds.) (2008). *Beyond Illustration: 2D and 3D Digital Technologies as Tools for Discovery in Archaeology*, BAR International Series 1805. Oxford: Archaeopress, 2008

Gillings, M. (2015). Mapping invisibility: GIS approaches to the analysis of hiding and seclusion. *Journal of Archaeological Science* **62**: 1–14

Hammond, N. and Tourtellot G. (1999). Shifting Axes: Spatial Expressions of Power at La Milpa. Paper presented at the 64th Annual Meeting, *Society for American Archaeology*. Chicago, IL. March 27th.

Hanks, W. (1990). *Referential Practice: Language and Lived Space among the Maya*. The University of Chicago Press, Chicago and London.

Houston, S., Stuart, D. and Taube, K. (2006). *The Memory of Bones: Body, Being, and Experience among the Classic Maya*. University of Texas, Austin.

King. J., Richards-Rissetto, H. and Landau K. (2015). Enter the Void: A GIS Analysis of the Visibility of Empty Spaces at Copan, Honduras. Paper presented at Society for American Archaeology 80th Annual Meeting, San Francisco, CA. April 2015.

Landau, K. (2015). Spatial Logic and Maya City Planning: The Case for Cosmology. *Cambridge Archaeological Journal* **25**(1): 275-92.

Lawrence, D. and Low, S. (1990). The Built Environment and Spatial Form. *Annual Review of Anthropology* **19**: 453-505.

Llobera, M. (2003). Extending GIS Based Analysis: The Concept of the Visualscape. *International Journal of Geographic Information Science* **1**(17): 25-48.

Llobera, M. (2006). What you see is what you get?: Visualscapes, visual genesis and hierarchy. In *Digital Archaeology: Bridging Method and Theory*, (Ed) P. Daly and T. Evans, Routledge, Taylor and Francis, New York and London, pp. 148-67.

Messenger, L. (1987). Community Organization of the Late Classic Southern Periphery of Mesoamerica: Expressions of Affinity. In *Interaction on the Southeast Mesoamerican Frontier: Prehistoric and Historic Honduras and El Salvador*, In E. J. Robinson (Ed), BAR International Series 327 (ii), pp. 385-420.

Muller, P., Vereenooghe, T., Wonka, P., Papp, I. and van Gool L. (2006). Procedural 3D reconstruction of Puuc buildings in Xkipché. *7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST*. M. Ioannides, D. Arnold, F. Niccolucci, and K. Mania (Ed).

Paliou, E. (2014). Visibility analysis in 3D built spaces: a new dimension to the understanding of social space. In *Spatial analysis and social spaces: Interdisciplinary approaches to the interpretation of prehistoric and historic built environments*, E. Paliou, U. Lieberwirth, and S. Polla (eds). Series: Topoi – Berlin Studies of the Ancient World 18.

Rapoport, A. (1990). *The Meaning of the Built Environment: A Nonverbal Communication Approach*. University of Arizona Press: Tucson.

Richards-Rissetto, H. (2010). *Exploring Social Interaction at the Ancient Maya City of Copán,*

Honduras: A Multi- Scalar Geographic Information Systems (GIS) Analysis of Access and Visibility. Unpublished PhD: University of New Mexico.

Richards-Rissetto, H. and Plessing, R. (2015). Procedural modeling for ancient Maya cityscapes: Initial methodological challenges and solutions. *Proceedings for Digital Heritage International Congress 2015,* Granada, Spain.

Richards-Rissetto, H. (2013). From mounds to maps to models: visualizing ancient architecture across landscapes. *Proceedings of Digital Heritage International Congress 2013*, Marseille, France.

Richards-Rissetto, H. and Landau, K. (2014). Movement as a means of social re(production): Using GIS to measure social integration in urban landscapes. *Journal of Archaeological Science* **41**: 365-75.

Saldaña, M. (2015). An Integrated Approach to the Procedural

Modeling of Ancient Cities and Buildings. *Digital Research in the Humanities* (print volume forthcoming).

Saldaña, M. and Johanson, C. (2013). Procedural Modeling for Rapid-Prototyping of Multiple Building Phases and Hypothetical Reconstructions of Early Rome. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XL-5/W1.

Thompson, A. and Prufer, K. (2015). Evaluating airborne LiDAR for detecting settlements and modified landscapes in disturbed tropical environments at Uxbenka, Belize. *Journal of Archaeological Science* **57**: 1-13.

von Schwerin, J., Richards-Rissetto, H., Remondino, F. and Agugiaro G. (2013). The MayaArch3D Project: A 3D WebGIS for Analyzing Ancient Maya Architecture and Landscapes at Copan, Honduras. *Literary and Linguistic Computing*. Oxford University Press.

Von Schwerin, J., Richards-Rissetto, H., Remondino, F., Grazia Spera, M., Auer, M., Billen, N., Loos, L. and Reindel M. (2016). Airborne LiDAR Acquisition, Post-Processing and Accuracy-Checking for a 3D WebGIS of Copan, Honduras. *Journal of Archaeological Science Reports*.

The Trajectories Tool: Amplifying Network Visualization Complexity

Alexandre Rigal
alexandre.rigal@epfl.ch
EPFL, Switzerland

Dario Rodighiero
dario.rodighiero@epfl.ch
EPFL, Switzerland

Loup Cellard
loup.cellard@epfl.ch
EPFL, Switzerland

Introduction

Network visualizations are the most complex visualizations possible, but sometimes they are not capable of describing system-complexity. Even if they are the most widely employed visualization techniques, they still have limitations. Indeed a) their relations are not sufficient to analyse complexity and b) networks do not distinguish between qualitative differences of represented entities.

Starting from the actual network model, how could one manipulate this visualization to improve complexity comprehension? In this paper, we propose a solution called *trajectory*. The trajectory has two major points of difference compared to the network: the trajectory a) represents not only distances, but also durations, and b) it displays kinetic entities according to their evolution with time.

The discourse is articulated around these four points. Considering that networks are tools widely used by digital humanists, we propose a new language to improve the quality of represented data: a new network based on a vertical timeline. Complexification of the network visualization is not just a new language, but also a tool that would give the field of Digital Humanities the most complex of all possible visualizations.

Networks

How could one improve the visualization of complexity? To answer this question, we need to investigate the qualities and defects of network visualizations. A network is a wonderful way to visualize complexity: in this model, limitless relations and entities can be mapped. A network neither draws frontiers nor imposes quantitative limits to relations and entities. How does one imagine an even more complex visualization? In other words, how does one draw a "greater infinity" of relations and entities? It seems impossible indeed; the network could possess interminable relations and entities. But a trick is still possible – to enrich entities by adding a temporal dimension.

Conventionally, network visualizations are created considering the interdependence between distance and attraction, i.e. the spatial relationship. In fact, network visualizations are stable images, keeping each entity immobile, without possible evolutions. Changes in connections and disconnections of entities happen because of temporal events taking place in reality. So, if we want to achieve an even more detailed representation of complexity, we need to introduce another dimension for relations and entities – the dimension of time.

Figure 1. Umberto Boccioni, *Spiral Expansion of Muscles in Action*, 191

Time and space are intimately linked through movement. To enhance the richness of the visual language, we need to visualize the movement of entities through static simulation, as done in *Spiral Expansion of Muscles in Action* by Umberto Boccioni, the Futurist artist who introduced the art of depicting sculpture in movement (Figure 1). We do not need to draw more actors or relations, it is enough to improve their representation with a more elaborated shape. In this way, complexity will be managed not only in terms of infinite spatial entities and relations, but also in terms of infinite time-based entities and relations.

Movement and complexity

If network visualization is to represent movement, it needs the fluidity through a new dimension of reality. This is the reason why relations and entities have to be both spatial and time-based. Movement lends a continuity to time and space, and is the key to better understand the representation. Consequently, the aim is to visualize the evolution of entities through a sequence of time-based networks.

Formerly, Kandinsky had noticed that nodes are fixed (1947, 32-35). To create dynamic visualizations, these nodes should progress. Now, network visualizations enriched by time could be said to be in motion. Visually speaking, network entities are represented by points; giving them a movement signifies that the points have to be represented through lines (Kandinsky 1947, 57; Ingold 2007). This solution allows us to lend more complexity to the basic point representation. Moreover, Kandinsky claims: “Considered in terms of substance, it [the point] equals zero” (1947, 25). The contrast between the point and the line in network visualizations finds another analogy in the Kandinsky’s thoughts, who considers the line as the antithesis to the point: “The line is, therefore, the greatest antithesis to the pictorial proto-element – the point” (Kandinsky, 1947: 57).

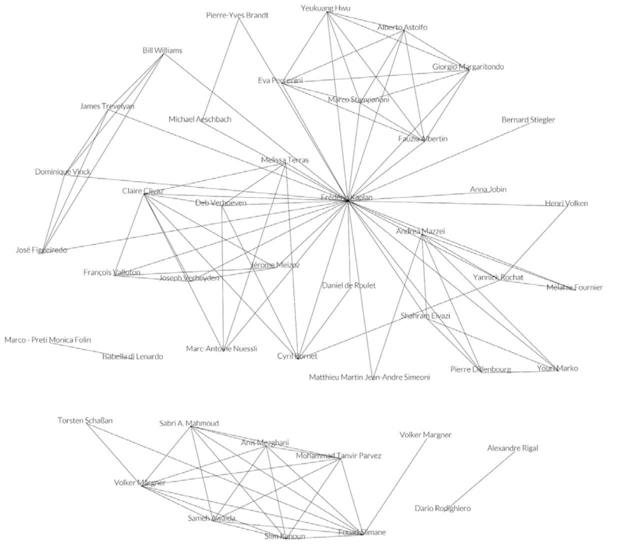


Figure 2. The DHLAB network is created by co-authoring: each node is an author and each edge a collaboration for a paper. Nodes represent also external collaborators

Thanks to the line representation, continuity enters the complexity of representation as the fabric stitching together the narrativity of several network sequences. Moreover, continuity implies an inherent relationship between both distance *and* duration. If the network visualization tends to be infinite in terms of spatial relations and entities, it is fairly weak to highlight time, another infinite dimension. The introduction of the trajectory brings another dimension to the visualization: the infinity of durations of relations and entities.

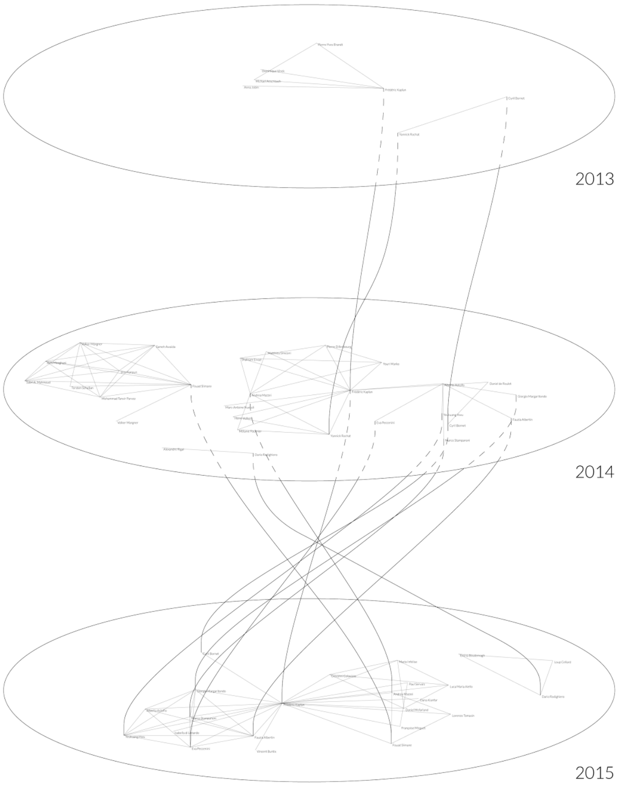


Figure 3. The DHLAB network is divided into years. Each trajectory represents an author who published during the years, tracing his continuity

The visualizations in practice

Data are extracted from Infoscience, EPFL’s publication repository. This repository is public, and everyone can access it and obtain the data presented in this paper. Technically speaking, we queried all publications associated with the DHLAB.

This article has two figures representing the DHLAB in Lausanne. Both visualizations present the same information in two different ways: the first shows collaborations on a plain surface – this figure is a classical network visualization (Figure 2); the second is based on the trajectory idea:

points previously arranged on a flat surface are transformed into lines reifying the continuity of people (Figure 3).

If we affirm that complexity representation has indeed been "complexified", how this could help digital humanists?

The first figure shows how the laboratory professor is placed at the center – the network would probably be very different without him and would be divided in parts. So, he is the core of the laboratory with good reason. The networks consists of two big clusters, which means that the professor is not included in all the publications of the laboratory (a rule that is compulsory in some labs).

The second figure immediately displays the literary production year by year. For example in 2013, the laboratory was relatively young to widely publish scientific works, collaborations have a greater size in 2014 and 2015. Compared to the network visualization, the trajectories depict the variability of links with time quite well by splitting the network into a sequence of networks.

The trajectories also describe the centrality of the professor better. For example in 2014, he basically worked with a group of people (shown at his left) and a colleague (at his right). In the same year, this colleague had a group of authors with whom he collaborated. Following the trajectories of that group, we can easily see that in 2014, it was not directly linked to the professor whereas in 2015, they all worked together. This is a detail revealed by trajectories that was not visible with classical network visualization.

The tool (conclusion)

Gyorgys Kepes wrote: "each new visual environment demands [...] new way of measuring" (1995, 13). This phrase introduces two concepts that would be of interest to the communities of Digital Humanities and Data Visualization: context and method. We could suppose that the context is given by humanists and the method by designers, but DH is the result of a blend between the two. Designers can not think of methods out of context, the opposite is true for humanists.

As Johanna Drucker says, we are experiencing "a momentary blindness among practitioners" (2011, 5). To overcome this blindness, we put together the efforts of designers and humanists: the visualization here described represents interdisciplinarity – the context and the method at the same time. For this reason, we think that trajectories are a good example. The DH community is close to having the sensibility to solve the interdisciplinarity issue.

For this reason, theory has to be combined with practice to produce solid research tools for the DH community. Trajectories are not simply a theoretical concept but also a tool that will be presented at the DH 2016 conference and which will be available to digital humanists for decomposing their networks, or, in other words, to lend representations their complexity.

Bibliography

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Boccioni, U. (1914). *Manifesto of Futurist Sculpture*.

Drucker, J. (2011). Humanities approaches to graphical display, *Digital Humanities Quarterly*, **5**(1): 1-21.

Holme, P. and Saramäki J. (2013). (Eds.) *Temporal Networks. Understanding Complex Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ingold, T. (2007). *Lines, A Brief History of Lines*, Routledge, London.

Lima, M. (2014). *The Book of Trees: Visualizing Branches of Knowledge*, Princeton Architectural Press, Princeton.

Kandinsky, W. (1947). *Point and Line to Plane*, The Solomon R. Guggenheim Foundation for the Museum of Non-Objective Painting, New York.

Kepes, G. (1995). *Langage of Vision*, Dover, New York.

Meeks, E. (2015). *D3.js in Action*. Shelter Island, NY: Manning.

Meirelles, I. (2013). *Design for Information*. Rockport Publishers.

Munster, A. (2013). *Anaesthesia of Networks: Conjunctive experience in Art and Technology*, MIT Press, Cambridge.

À la Croisée des Discours Littéraire et Scientifique : La Comparaison comme Haute Figure Dialogique

Marine Riguet
marineriguet@gmail.com
Université Paris-Sorbonne, France

Suzanne Mpouli
mpouli@acasa.lip6.fr
Université Pierre et Marie Curie, France

Introduction

S'intéresser à la comparaison, à l'échelle du discours, c'est interroger sa fonction d'amarrage, de point charnière, entre plusieurs champs énonciatifs distincts ; c'est questionner la façon dont deux discours entrent en interaction et s'activent l'un l'autre sur le plan structurel, à travers un nouveau réseau de ressemblances/différenciations qui participe de leur affirmation singulière.

Cette étude exploratoire se concentre sur le dialogue particulièrement riche qu'entretiennent les discours scientifique et littéraire entre 1860 et 1928, et plus précisément sur la convocation, au sein de la critique littéraire, de tout un champ notionnel emprunté aux sciences dites exactes (biologie, physique, chimie, botanique…). À une époque où émergent la biologie moderne, la médecine expérimentale, les principes de thermodynamique, ou encore l'évolutionnisme de Darwin et Spencer, le discours scientifique ne se cantonne plus à une sphère dévolue aux spécialistes, mais nourrit au contraire des mutations sociales, éthiques et esthétiques, au travers desquelles la littérature redéfinit sa place.

Au vu de la dimension analogique de la comparaison, on pourrait supposer qu'elle tient un rôle central dans cet appareillement des discours littéraire et scientifique, captant et catalysant en quelque sorte le « dialogisme discursif » (Adam, 2005) au niveau microscopique. Il s'agit donc de déterminer ses différentes fonctions et manifestations au cœur de cette intertextualité qu'elle crée.

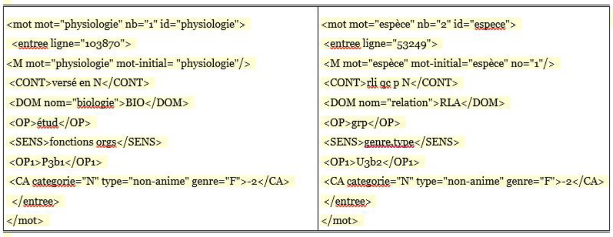
Méthode automatisée

Le besoin d'une méthode automatisée et le recours à des outils informatiques nous ont paru incontournables afin de pouvoir traiter un corpus à la fois signifiant et vaste. Nous avons donc cherché à identifier et à extraire automatiquement les comparaisons concernées au sein d'un corpus numérisé homogène, comprenant 49 auteurs pour un total de 140 ouvrages de critique littéraire française. Pour ce faire, deux outils ont été utilisés :

- un algorithme permettant d'identifier des comparaisons potentielles construites autour de différents types de marqueurs de la comparaison ainsi que les termes qu'elles mettent en parallèle (Mpouli & Ganascia, 2015) ;

- le Dictionnaire électronique des mots (Dubois & Dubois-Charlier, 2010), un dictionnaire au format XML qui fournit pour chaque entrée un champ lexical et une sous-classe sémantique (cf. Figure 1), offrant ainsi la possibilité de ne retenir que les comparaisons dans lesquelles le comparant ou le comparé relève exclusivement du champ lexical de l'une des sciences exactes prédéfinies (chimie, chirurgie, médecine, physiologie, physique, mathématiques, astronomie, biologie, géologie, zoologie et botanique).

Précisons cependant que le terme « espèce », qui occupe une place importante dans le vocabulaire scientifique de l'époque, a dû être ajouté manuellement car le dictionnaire ne le plaçait dans aucun des domaines scientifiques concernés.



```
<mot mot="physiologie" nb="1" id="physiologie">
 <entree ligne="103870">
<M mot="physiologie" mot-initial= "physiologie"/>
 <CONT>versé en N</CONT>
 <DOM nom="biologie">BIO</DOM>
 <OP>étud</OP>
 <SENS>fonctions orgs</SENS>
 <OP1>P3b1</OP1>
 <CA categorie="N" type="non-anime" genre="F">-2</CA>
 </entree>
</mot>
```

```
<mot mot="espèce" nb="2" id="espece">
 <entree ligne="53249">
 <M mot="espèce" mot-initial="espèce" no="1"/>
 <CONT>rli qc p N</CONT>
 <DOM nom="relation">RLA</DOM>
 <OP>grp</OP>
 <SENS>genre,type</SENS>
 <OP1>U3b2</OP1>
 <CA categorie="N" type="non-anime" genre="F">-2</CA>
 </entree>
</mot>
```

Figure 1. Exemple de deux entrées dans le Dictionnaire électronique des mots

Les termes ainsi recueillis peuvent être regroupés en trois grands ensembles : les disciplines scientifiques, les métiers et les concepts qui leur sont rattachés. Si on ne considère que les termes scientifiques utilisés comme comparant, on se rend compte que le domaine de la zoologie, contenant entre autres les termes « espèce » et « animal », est le plus fortement représenté (42 % des occurrences). La fréquence de ces comparants permet en outre de dégager des groupes d'auteurs bien distincts (cf. Figure 2).

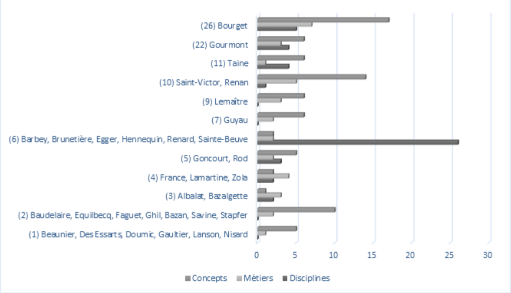


Figure 2. Répartition des auteurs en fonction de leur fréquence d'utilisation de termes scientifiques comme comparant. Seuls ont été considérés les concepts comptant au moins 5 occurrences dans le corpus

Activation d'un dialogue interdisciplinaire

De fait, une lecture plus poussée des résultats confirme deux positionnements antagonistes de la critique face au discours scientifique. D'une part, tout un pan de la critique dénigre l'aridité ou l'arrogance d'un discours scientifique qui permet, par contraste, de valoriser une vision romantique et symboliste de la littérature. La comparaison introduit en ce sens un rapport de supériorité au bénéfice du discours littéraire.

Livre	Comparaison extraite	Marqueur	Comparé	Comparant
P. de Saint-Victor, Le Théâtre contemporain.	Depuis quelques années déjà, l'auteur de l'Ami des femmes exerce la morale comme une chirurgie ; il lui prête l'impudeur tranchante d'une science expérimentale qui a le droit de tout éventrer et de tout décrire.	Comme	morale	chirurgie
A. de Lamartine, Cours familier de littérature, t. 5.	[Les psychologues] n'arrivent qu'à s'embrouiller dans leurs définitions, à se contredire dans leurs distinctions, à se perdre dans leur analyse; et, comme les chimistes, leurs émules, quand ils veulent retirer de leur creuset les principes de l'âme humaine et dire : La voilà ! ils ne tiennent sous leur plume ou sous leurs doigts qu'une pincée de cendre…	comme	[psychologues]	chimistes

De l'autre, une critique littéraire dite « scientifique », regroupant entre autres Sainte-Beuve, Brunetière, Guyau, Bourget, Taine, Renan, Hennequin, Lemaître, Goncourt, Zola ou Renard, instaure au contraire un rapport d'égalité avec un discours scientifique apte à enrichir le discours littéraire.

J.-M. Guyau L'Art au point de vue sociologique.	Dans la composition des caractères, par exemple, l'art combine, comme les chimistes dans la synthèse des corps, des éléments empruntés à la réalité.	Comme	art.	chimistes
E. Zola cité par L. Bazalgette, L'Esprit nouveau dans la vie artistique, sociale et religieuse.	Nous devons opérer sur les caractères, sur les passions, sur les faits humains et sociaux, comme le chimiste et le physicien opèrent sur les corps bruts, comme le physiologiste opère sur les corps vivants.	Comme	nous	chimiste, physicien, physiologiste

Au sein de ce second type de critique, la comparaison apparaît comme une figure de premier plan par son aptitude à cimenter un rapport de type analogique entre les discours.

Le système analogique d'une critique littéraire « scientifique »

En rhétorique, la comparaison contribue souvent, de par sa fonction évocatrice, à donner plus de poids à une argumentation. Si plusieurs usages de la comparaison peuvent ici être analysés, tous semblent néanmoins obéir à une même stratégie discursive : l'image provoquée par le rapprochement de l'objet littéraire avec un élément exogène, puisé dans le domaine scientifique, vient servir le discours critique sur un plan rhétorique immédiat. Agrandir le champ notionnel par un rapprochement analogique, c'est par là-même généraliser la portée d'un énoncé qui devient exportable à d'autres domaines. En ce sens, la comparaison est mise au service d'un processus de valorisation, dans la mesure où le discours littéraire, trouvant un pied d'égalité avec le discours scientifique, reçoit simultanément une légitimité qui lui est extérieure – sans compter la forte caution institutionnelle dont bénéficie à cette époque le discours scientifique.

Cette mise en place d'un système analogique peut ainsi s'observer autour de « concepts nomades » (Stengers, 1987) que la comparaison rend nettement identifiables, telle que la notion d'espèce, dont la circulation et la réappropriation par le discours littéraire sont ici particulièrement remarquables.

P. Bourget Études et Portraits, t. I.	Des espèces littéraires existent, analogues aux espèces vivantes, constituées par des caractères propres et irréductibles les unes aux autres, malgré l'unité de composition de notre monde intellectuel.	analogue à	Espèces littéraires	Espèces vivantes

Cette égalité introduite par la comparaison entre domaines littéraire et scientifique n'est jamais stricte, mais structurelle, comme le montre les exemples repris ci-dessous. La comparaison pose donc une ressemblance entre des rapports, de sorte que la critique littéraire peut traiter la littérature comme la science traite son objet : la littérature, l'art, deviennent par la force de l'analogie champs de savoir. En servant de « modèle analogue » (Black, 1962), non seulement la science légitime le projet de la critique littéraire, mais elle opère consécutivement un renversement axiologique en imposant la valeur de vérité, à laquelle les valeurs esthétiques et éthiques se retrouvent subordonnées.

E. Renan, Discours et conférences.	Car il y a une logique dans une tragédie en cinq actes comme dans un mémoire de physiologie, et la règle des ouvrages de l'esprit est toujours la même : être égal à la vérité, ne pas l'affaiblir en s'y mêlant, se mettre tout entier à son service, s'immoler à elle pour la montrer seule, dans sa haute et sereine beauté.	Comme	tragedie	mémoire de physiologie

A. Albalat Le Mal d'écrire et le roman contemporain.	L'étude comparative des procédés de composition dans les œuvres écrites démontre au contraire qu'il y a en art des lois d'examen invariables, des vérités esthétiques démontrables comme les mathématiques, des principes fixes, une évidence positive et irrésistible, une certitude enfin de filiation et de descendance commune à toutes les écoles et élucidant toutes les productions.	Comme	vérités	mathématiques

Enfin, la comparaison, autant qu'elle rassemble, se fait lieu de confrontation entre les discours. L'identité même du discours se voit ainsi mise en jeu, dans la mesure où le rapprochement analogique sert aussi, pour la critique littéraire, de mise en rivalité avec un discours scientifique dont elle cherche à la fois à emprunter des structures et à rejeter l'ascendant. Par la comparaison, qui contrairement à la métaphore ne « fusionne » pas, le dialogisme discursif participe finalement de la construction identitaire du discours littéraire, dans ce sens où « l'Autre n'est pas seulement la contrepartie du Même, mais appartient à la constitution intime de son sens » (Ricœur, 1990). En ce sens, on pourrait se demander dans quelle mesure et sous quelle forme la métaphore contribue elle aussi à la construction de ce nouveau discours critique.

Bibliography

Adam, J.-M. (2005). *La Linguistique textuelle. Introduction à l'analyse textuelle des discours.* Paris: Armand Colin.

Black, M. (1962). *Models and metaphors. Studies in Language and Philosophy*. Ithaca: Cornell University Press.

Dubois, J. and Dubois-Charlier, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration, *Langages*, **3**(179-180): 31-56.

Mpouli, S. and Ganascia, J.-G. (2015). Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives. *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, Caen.

Ricœur, P. (1990). *Soi-même comme un autre.* Paris: Seuil.

Stengers, I. (ed). (1987). *D'une science à l'autre. Des concepts nomades.* Paris: Seuil.

Modelling Music Reception: An Ontology For Representing Interpretations of Richard Wagner's Leitmotifs

Carolin Rindfleisch
carolin.rindfleisch@music.ox.ac.uk
University of Oxford, United Kingdom

1. Background

Richard Wagner's leitmotifs (Whittall, 2003), the compositional technique and its aesthetic implications as well as the individual motives, their musical characteristics, meanings, and transformations through the works, have been subjected to a wide variety of interpretations throughout different cultural and historical situations. The popular Wagner discourse in particular – in work introductions, guidebooks, programme notes, commentaries and motivic annotations in libretti and piano scores – focused on the extraction, description and presentation of leitmotifs in order to convey their interpretation of Wagner's work. Thus, the phenomenon 'leitmotif' cannot only be defined as a compositional technique, but also as a reception practice (Grey, 1996; Rümenapp, 2002; Thorau, 2003). This study aims at a large-scale comparison of these reception practices that reaches across cultures and time periods and is able to not only describe general characteristics of sources, but to compare interpretations down to conceptions of individual motives. Semantic Web technologies (Berners-Lee et al., 2001; Bizer et al., 2011), with the enhanced capabilities of linking, searching and structuring of data they provide, may facilitate this undertaking. This paper describes an ontology that models the structure and content of leitmotif interpretations, which shall enable the semantic annotation of sources and the structured representation of the interpretations they contain. A particular focus lies thereby on the methodological considerations and design decisions that stand behind it.

2. Methodological Considerations

Semantic Web Research and Musicology are two distinct domains, each with their own conventions, terminologies, and patterns of thought and discourse. Attempts to integrate them thus have to deal with the following questions: "How do [musicological] methods translate into methods of Semantic Web Research?" (Kummer, 2011) Can we conceive of a musicological ontology, which not only forms the basis to represent discourse in novel ways, but also works as an analytical device? How can we take advantage of the commonalities and accommodate the differences between methodologies and ways of thinking, such as different approaches to the presentation of knowledge and the construction of complexity, and the confrontation of different values of flexibility, ambiguity, and metaphoricity on the one hand and explicit definitions and taxonomies on the other? There are several key strategies by which the described ontology aims to achieve compliance with both domains: 1) a historical, critical and philosophical 'awareness' within the terminology and its specifications, as well as the connection to a shared vocabulary in the form of musicological encyclopaedias; 2) the incorporation of flexibility, ambiguity and nuance into the explicit model; 3) a direct correspondence between the ontology and both the document and "thought structure" of the interpretations it models.

3. Related Work

Existing approaches that apply Semantic Web technologies to music or musicology are mostly concerned with organising and retrieving musical data. The Music Ontology aims at modelling the "music production workflow", and focuses on managing "music-related data" of recorded and foremost popular music (Raimond, 2007). Since this domain specification also shapes the terminology and definitions it employs, its application in the scholarly context of music history can be problematic: it does not comply with the terminological standards and conventions of musicology, and does not provide sufficient concepts to describe the variety of data that contextualises music and that is vital for music-historical study. The "musicSpace" project (Bretherton et al., 2009) aimed at enhancing the possibilities of musicological research by "integrating access to many of musicology's leading data sources", gathering factual information from this data and metadata and making this information accessible and searchable via a single interface. While sharing common goals and objectives, the project at hand aims at going beyond factual information and at analysing the structure and content of a specific type of musicological discourse, modelling assignments people attribute to a musical work throughout the history of its interpretation. Related approaches can also be found in different disciplines, such as Cultural Heritage and Museology (Doerr, 2003): Traditional approaches in musicology tended to treat works as self-contained conceptual objects; therefore, the reception or interpretation history of a musical work can, to a certain extent, be seen in correspondence to the provenance of an object. Relevant are furthermore ontologies that aim at modelling narratives or, more general, the content or structure of sources in different contexts, such as the OntoMedia ontology (Jewell et al., 2005), Story Fountain (Mulholland et al., 2004), or Curate (Mulholland et al, 2012).

4. Structure and Design

Analyses and interpretations of musical works, even though their content, analytical approach etc. might be completely different, on a basic level often rely on similar structures and constituent elements. These commonalities increase within the "specialised" analytic or interpretative discourse that is concerned with, for instance, a particular composer, a particular form, style or genre, which shares patterns and conventions of discourse and a specified vocabulary. This ontology, therefore, aims to model leitmotif interpretation as a special case of musicological discourse: its design is based on an analysis of representative source documents, and its structure and scope is informed by the structure and content of these interpretations and their presentation principles.

The ontology negotiates between several layers that model different dimensions. The first contains bibliographical metadata on the source documents, while the second concerns the source type and the corresponding internal document structure. The third level comprises the content structure, and is most significant for the comparisons of the different leitmotif interpretations.

Guidebooks, such as Hans von Wolzogen's *Thematischer Leitfaden* (Wolzogen, 1876), are already highly structured documents: They follow the chronology of the work, and "narrate" the music by extracting leitmotifs and describing their appearances, interactions and transformations through the course of the work. This interpretational concept of a leitmotif is, however, not the same as the compositional concept: they are integrated, self-contained constructs, abstractions from the musical context that tie together a musical phenomenon and its interpretation. They are represented as a collection of several constituent elements, such as a notation example fixing the motive's shape, and a descriptive name defining its semantic reference (see figure 1). Which constituent elements exactly form a leitmotif concept differs between sources, and is characteristic for the source type or the interpretation approach. However, whereas within one guide, leitmotif concepts are stable and easily identifiable, the comparison between a large number of sources undermines permanent constructions: the same name can be found applied to different musical phenomena, or the same musical shape can be assigned different names; and even seemingly small nuances can point to significant differences in interpretations and aesthetic attitudes: the names 'Schicksalsmotiv'

(Wolzogen, 1876) and 'Schicksalsfrage' (Porges, 1882), for instance, imply a completely different relationship between the name and the musical phenomenon: while the former establishes a denoting connection, with the motive acting as a sign for its semantic reference, the second establishes a metaphoric connection between the musical structure and the name, as the rhetoric characteristics of the 'Frage' ('question') parallel the musical characteristics of the motive. The varied combinations of constituent elements in different sources further complicate a comparison: how can we compare leitmotif concepts that don't share any features? The ontology applies several strategies in order to cope with these difficulties: 1) Rather than providing a fixed definition of an interpretational leitmotif concept and its structure, the notion of a leitmotif concept is defined through its relationship to its constituent elements. Thus, the ontology does not prescribe a particular structure in which the actual instances have to be fitted, but stays flexible and can accommodate varying interpretational approaches. 2) The ontology introduces three meta-constituents, that are assigned in the analytic process and that shall introduce a common basis for comparison: the reference to a point in the work, the reference to a semantic sphere, and the reference to a basic musical shape (see figure 1). 3) Within these three categories, there is a set of relationships that aims at expressing a differentiated spectrum of influence and independence, identity and contrast, similarity and difference. In addition to the described structural features, the ontology also introduces different levels of abstraction, so that leitmotif interpretation is modelled as a special case of musicological discourse in general and the ontology remains potentially extensible to accommodate other forms of music interpretation and discourse.

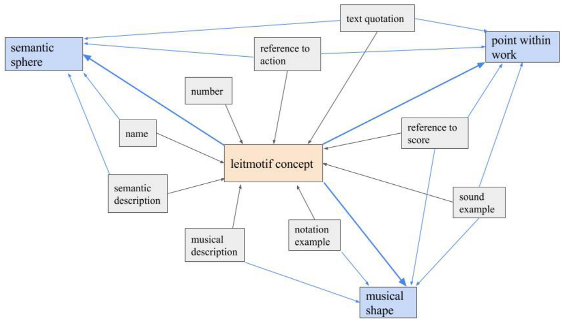


Figure: Possible constituents of a leitmotif concept

5. Conclusion

The described ontology forms the basis of a semantic annotation of the digitised sources, which shall facilitate the systematic comparison of a wide variety of leitmotif interpretations and enable to answer questions such as the following: Which musical shapes have been associated with a certain semantic sphere? How has a certain set of musical characteristics been classified in different interpretations and reception contexts? Being able to reorganise and query the reservoir of leitmotif concepts according to such questions can provide insights about larger contexts, patterns and constellations and form the starting point of more detailed investigations. Furthermore, beyond this particular analysis, it forms a first step towards a new structure of musicological discourse, as it allows to link elements of musical analyses, interpretations or reception documents together in novel ways.

Bibliography

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, **284**(5): 34-43.

Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked Data: The Story so far. *Semantic Services, Interoperability and Web Applications. Emerging Concepts*, pp. 205-27.

Bretherton, D. et al. (2009). Integrating musicology's heterogeneous data sources for better exploration. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pp. 27-32.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, **24**(3): 75-92.

Grey, T. (1996). ...wie ein rother Faden. On the origins of 'leitmotif' as critical construct and musical practice. In Bent, I. (Ed), *Music theory in the age of Romanticism*. Cambridge: Cambridge University Press, pp. 187-210.

Jewell, M.O. et al. (2005). OntoMedia: An ontology for the representation of heterogeneous media. In *Proceeding of SIGIR workshop on Mutlimedia Information Retrieval.*

Kummer, R. (2011). Semantic Technologies for Manuscript Descriptions – Concepts and Visions. *Codicology and Palaeography in the Digital Age 2. Schriften des Instituts für Dokumentologie und Editorik, 3*. Norderstedt, pp. 133-55.

Mulholland, P., Collins, T. and Zdrahal, Z. (2004). Story fountain: intelligent support for story research and exploration. *Proceedings of the 9th international conference on Intelligent user interfaces*, pp. 62-69.

Mulholland, P., Wolff, A. and Collins, T. (2012). Curate and storyspace: an ontology and web-based environment for describing curatorial narratives. *The Semantic Web: Research and Applications.* Berlin: Springer, pp. 748-62.

Porges, H. (1882). *Die Bühnenproben zu den Bayreuther Festspielen des Jahres 1876*, Vol. **2**, Die Walküre, Chemnitz: Schmeitzner.

Raimond, Y., et al. (2007). The Music Ontology. *Proceedings of the International Conference on Music Information Retrieval (ISMIR).*

Rümenapp, P. (2002). *Zur Rezeption der Leitmotivtechnik Richard Wagners im 19. Jahrhundert*. Wilhelmshaven: Noetzel.

Thorau, C. (2003). *Semantisierte Sinnlichkeit. Studien zu Rezeption und Zeichenstruktur der Leitmotivtechnik Richard Wagners*. Stuttgart: Franz Steiner Verlag.

Whittall, A. (2003). Leitmotif. *The new Grove dictionary of music online.*

Wolzogen, H. V. (1876). *Thematischer Leitfaden durch die Musik zu Richard Wagners Festspiel Der Ring des Nibelungen,* Leipzig: Schloemp.

Digging into ECCO: Identifying Commonplaces and other Forms of Text Reuse at Scale

Glenn Roe
glenn.roe@anu.edu.au
Centre for Digital Humanities Research, Australian National University, Australia

Clovis Gladstone
clovisgladstone@uchicago.edu
The ARTFL Project, University of Chicago, USA

Robert Morrissey
rmorriss@uchicago.edu
The ARTFL Project, University of Chicago, USA

Mark Olsen
markymaypo57@gmail.com
The ARTFL Project, University of Chicago, USA

Background

Commonplaces are a particular instance of historical text reuse (Dacome, 2004; Allan 2010; Blair, 2011). This paper describes our efforts at identifying commonplaces in the Gale-Cengage *Eighteenth Century Collections Online* (ECCO) database. Given the size of this collection, as well as the state of the data in terms of its OCR output, identifying shared passages that exhibit the textual characteristics of commonplaces – e.g., are relatively short, repeated, and rhetorically significant – is a non-trivial computational task. In our previous work on text reuse, we came across numerous examples of textual borrowings and shared passages that we considered possible commonplaces (Allen et al., 2010; Horton et al., 2010; Roe, 2012). We expanded this work into a Digging into Data Round 3 project using similar methods to explore the more than 200,000 works contained in ECCO, a dataset that represents most of the printed literary and scientific output in Britain from 1700 to 1799[1].

Previously we developed a sequence alignment algorithm for the identification of large-scale text reuse[2]. This algorithm, called PhiloLine, generates a list of similar passages (based on a set of flexible matching parameters) shared between any two texts. This simple approach allows us to find borrowings and other instances of text reuse, from quotations to uncited passages and paraphrases, over large heterogeneous corpora (Edelstein et al., 2013). Historical text reuse detection is a burgeoning field within the digital humanities, whether focussed on literary allusion (Coffee et al., 2012; Coffee et al., 2014), paraphrase (Büchler et al., 2011), influence (Büchler et al., 2014), or networks of reprinting (Smith et al., 2015). While all these projects address text reuse in slightly different ways, the flexibility and scalability offered by PhiloLine, coupled with our familiarity with the system, offered significant advantages over other approaches. We thus aimed to use PhiloLine to compare the ECCO corpus to itself, compile a list of the most frequent shared passages, and from there evaluate these passages in order to build a database of potential commonplaces.

Eliminating duplicates

The scope and scale of the ECCO dataset represented a major hurdle both in terms of computational expense and evaluation of the matching algorithm. Faced with more than 32 million pages of text, any manipulation of the data takes on significant proportions. To put this in perspective, comparing ECCO's 200,000 documents to each other means making some 40 billion pairwise comparisons and then storing and evaluating the output. Fortunately, our focus on commonplaces requires us to dramatically reduce the number of comparisons. We needed, for instance, to eliminate duplicate or near-similar texts in order to reduce the number of documents for comparison. The most obvious method would be to compare all the words in each work, and define a similarity threshold beyond which we consider two works to be the same. But, given the unequal quality of the OCR in the ECCO dataset, the reliability of any algorithm meant to detect similarity between two texts is very low. Two identical texts, for instance, can potentially only share 20% of individual tokens due to the quality of the OCR. As a result, we decided to focus our efforts on comparing document metadata instead, as it is of excellent quality.

Our methodology consisted in comparing titles in the dataset using a cosine similarity algorithm (Singhal, 2011). For our purposes, we determined a minimal similarity index to automatically determine whether two texts were the same, that is to say a re-edition of the same work. Beyond a certain threshold score, the newest document in terms of date of publication is automatically flagged as a duplicate. If it so happens that the minimal score is not reached, but still remains high, we compare authors, and if these are the same, we similarly flag the most recent document as a duplicate:

```
Psuedocode for title matching algorithm

if score >= 0.8:
   return True
elif score > 0.6 and score < 0.8:
   if author == other_author:
       return True
   else:
       return False
```

The document with the oldest publication date serves as the "source" text. Using this method, we were able to reduce the size of the corpus by 43%, eliminating 88,850 documents. We were thus left with 116,700 unique texts on which to run our matching algorithm.

Detecting similar passages

Similar passage detection requires a one-to-one document comparison. Rather than compare all 116,700 texts to one another, we decided to leverage Gale's thematic divisions, and limit the comparison task to individual modules[3]. While PhiloLine's matching capacity is powerful, it is necessary to understand its underlying algorithm in order to configure the tool properly. Briefly, PhiloLine's operational logic is to compare sequences of words, or n-grams, and determine the presence of a shared passage according to the number of common contiguous n-grams between two sequences. For example, the following text from Shakespeare's *The Tempest* is rendered by as a set of overlapping n-grams (where n = 3):

The cloud-capped towers, the gorgeous palaces,
The solemn temples, the great globe itself—
Yea, all which it inherit—shall dissolve
- Shakespeare, *The Tempest*, Act 4, Scene 1 ca. 1611

```
Trigrams: cloud_capped_towers, capped_towers_gorgeous, towers_gorgeous_palaces,
gorgeous_palaces_solemn, palaces_solemn_temples, solemn_temples_globe,
temples_globe_itself, globe_itself_yea, itself_yea_inherit, yea_inherit_shall,
inherit_shall_dissolve
```

Trigram generation and stopword removal are thus the main parameters we apply to transform texts prior to the sequence alignment process. Once this is done, we proceed with the text-sequence comparisons. Below is an example of just such an alignment of sequences drawn from the Literature and Language module:

an Hour of virtuous Liberty, Is worth a whole Eternity in Bondage
- Joseph Addison

```
Trigrams: hour_virtuous_liberty, virtuous_liberty_eternity, liberty_eternity_bondage
```

an hour, of virtuous liberty Is worth a whole eternity in bondage
- James Thomson

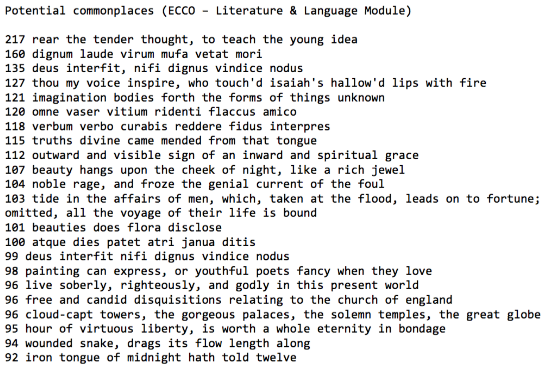
```
Trigrams: hour_virtuous_liberty, virtuous_liberty_eternity, liberty_eternity_bondage
```

In this case, we note the perfect alignment, which PhiloLine detected because there are at least three contiguous trigrams in common between both passages.

Using these base parameters (overlapping trigrams with stopwords removed), we compared the ECCO corpus to itself on a module-by-module basis. The output of this comparison ranged from 3.5 million common passages in the Literature and Language module, to almost 17 million possible commonplaces in Religion and Philosophy. Identifying these common passages is thus only a first step. Even after significant duplicate reduction, the sheer scale of the passages that require post-processing evaluation is daunting.

From similar passages to commonplaces

To attack this problem, we treat commonplaces generically as the repeated use of the same passage - more or less similar - in a minimum number of different authors. We began by grouping all source passages that were identical in order of frequency. Given that commonplaces are normally short expressions, at the most no longer than several sentences, we restricted our search to passages containing a minimum of five words and a maximum of 75:



```
Potential commonplaces (ECCO - Literature & Language Module)

217 rear the tender thought, to teach the young idea
160 dignum laude virum mufa vetat mori
135 deus interfit, nifi dignus vindice nodus
127 thou my voice inspire, who touch'd isaiah's hallow'd lips with fire
121 imagination bodies forth the forms of things unknown
120 omne vaser vitium ridenti flaccus amico
118 verbum verbo curabis reddere fidus interpres
115 truths divine came mended from that tongue
112 outward and visible sign of an inward and spiritual grace
107 beauty hangs upon the cheek of night, like a rich jewel
104 noble rage, and froze the genial current of the foul
103 tide in the affairs of men, which, taken at the flood, leads on to fortune;
omitted, all the voyage of their life is bound
101 beauties does flora disclose
100 atque dies patet atri janua ditis
99 deus interfit nifi dignus vindice nodus
98 painting can express, or youthful poets fancy when they love
96 live soberly, righteously, and godly in this present world
96 free and candid disquisitions relating to the church of england
96 cloud-capt towers, the gorgeous palaces, the solemn temples, the great globe
95 hour of virtuous liberty, is worth a whole eternity in bondage
94 wounded snake, drags its flow length along
92 iron tongue of midnight hath told twelve
```

A cursory glance at this list reveals several variants of the same passage that need to be merged in order to better represent a single commonplace. If we take the following passage from the Scottish poet James Thomson, for instance:

Then infant reason grows apace, and calls For the kind hand of an assiduous care. Delightful talk! to rear the tender thought, To teach the young idea how to shoot, To pour the freft infiruAion o'er the mind, 1150 To breathe enlivening spirit, and to fix The generous purpose in the glowing breast.

We notice that the reuse of this passage in other authors can vary significantly.

Gentleman of the Middle Temple (1775):

How glorious would her matron employments be, to hear the tender thought, to teach the young idea how to Jhoot; to be at once the precept and example to her family of every thing that was good, every thing that was virtuous.

Mrs Lovechild (1790):

Happy the Mother "Distilling knowledge through the lips of "love !"- ' Delightful talk! to rear the tender thought, "To teach the young idea how to shoot", To pour the fresh inltrution o'er the mind !'Lines which will never cease to be quoted...

Given the variability in the reuse of any given passage, as well as the approximate quality of the OCR, we developed a new algorithm that could match similar passages in a way that was both precise, and yet more flexible than PhiloLine. The algorithm uses the same n-grams as PhiloLine, though they are constructed differently. Rather than use overlapping trigrams, as we do for sequence matching, here we use alternating bigrams for increased flexibility:

Then infant reason grows apace, and calls For the kind hand of an assiduous care.

```
Bigrams: infant_grows, reason_apace, grows_calls, apace_kind, calls_hand,
kind_assiduous, hand_care
```

By skipping a word in the creation of these bigrams, we create n-grams that are both rarer than in-sequence bigrams, but also more common than in-sequence trigrams. In essence, these bigrams are flexible trigrams where the middle word is ignored. In this manner we can alleviate some of the issues that come from the dirty OCR. As there is a higher probability for a regular trigram to contain a wrongly identified letter, it has a higher chance of being unique, therefore making it less reliable for similarity matching than our flexible trigrams.

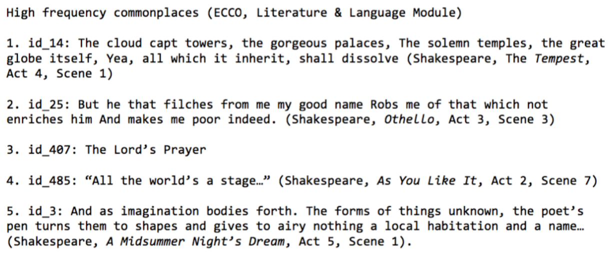
Finally, we needed to take into account the different lengths of these passages, as some are much longer than others. This variability led to the introduction of a coefficient that accounts for varying lengths, and allows us to automatically determine the minimum number of matching n-grams needed to establish similarity between two passages. For instance, if two passages of 30 words must share 4 bigrams, a passage of 30 words and another of 50 should share more bigrams to retain the same level of similarity.

Using the above methods, we were able to merge various uses of a single source passage and assign them a unique identifier. We can therefore now identify the highest frequency commonplaces in the Language and Literature module (see Appendix 1), as well as the most 'highly commonplaced' authors, i.e., those that generated the most shared passages (see Appendix 2). These preliminary results suggest that commonplacing, as an intertextual reading/writing practice, was alive and well in 18th-century England. Digging into a dataset such as ECCO can thus offer us new perspectives from which to view and understand 18th-century print culture, provided we unearth more than we cover up.

Future Work

We aim to release an interactive database of possible commonplaces in early 2016. The database will allow users to navigate the ECCO dataset via the commonplaces, most commonly cited authors and works, and visualize commonplace use and practices over time. We will also introduce several curated datasets that pre-date the 18th century, and that can act as a control on sources that fall outside the date boundaries of our data. These possible datasets include the King James Bible, Classical Latin texts, and EEBO-TCP[4]. Further goals for this project include merging the module-specific results into one large pool of potential commonplaces that reach across disciplinary boundaries; developing a user interface that allows for commonplace curation as a form of crowdsourcing; and introducing non-English datasets for comparison in order to find instances of multi-lingual commonplace practices.

Appendix 1



High frequency commonplaces (ECCO, Literature & Language Module)

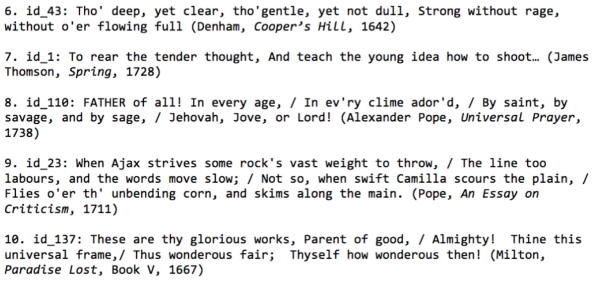
1. id_14: The cloud capt towers, the gorgeous palaces, The solemn temples, the great globe itself, Yea, all which it inherit, shall dissolve (Shakespeare, The *Tempest*, Act 4, Scene 1)

2. id_25: But he that filches from me my good name Robs me of that which not enriches him And makes me poor indeed. (Shakespeare, *Othello*, Act 3, Scene 3)

3. id_407: The Lord's Prayer

4. id_485: "All the world's a stage…" (Shakespeare, *As You Like It*, Act 2, Scene 7)

5. id_3: And as imagination bodies forth. The forms of things unknown, the poet's pen turns them to shapes and gives to airy nothing a local habitation and a name… (Shakespeare, *A Midsummer Night's Dream*, Act 5, Scene 1).



6. id_43: Tho' deep, yet clear, tho'gentle, yet not dull, Strong without rage, without o'er flowing full (Denham, *Cooper's Hill*, 1642)

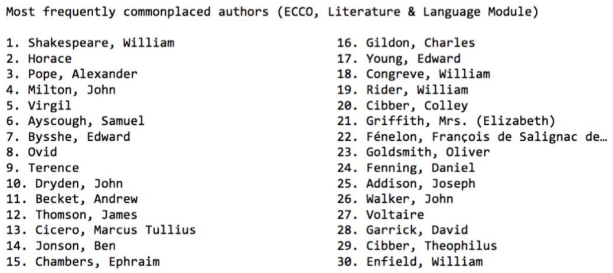
7. id_1: To rear the tender thought, And teach the young idea how to shoot… (James Thomson, *Spring*, 1728)

8. id_110: FATHER of all! In every age, / In ev'ry clime ador'd, / By saint, by savage, and by sage, / Jehovah, Jove, or Lord! (Alexander Pope, *Universal Prayer*, 1738)

9. id_23: When Ajax strives some rock's vast weight to throw, / The line too labours, and the words move slow; / Not so, when swift Camilla scours the plain, / Flies o'er th' unbending corn, and skims along the main. (Pope, *An Essay on Criticism*, 1711)

10. id_137: These are thy glorious works, Parent of good, / Almighty! Thine this universal frame,/ Thus wonderous fair; Thyself how wonderous then! (Milton, *Paradise Lost*, Book V, 1667)

Appendix 2



Most frequently commonplaced authors (ECCO, Literature & Language Module)

1. Shakespeare, William
2. Horace
3. Pope, Alexander
4. Milton, John
5. Virgil
6. Ayscough, Samuel
7. Bysshe, Edward
8. Ovid
9. Terence
10. Dryden, John
11. Becket, Andrew
12. Thomson, James
13. Cicero, Marcus Tullius
14. Jonson, Ben
15. Chambers, Ephraim
16. Gildon, Charles
17. Young, Edward
18. Congreve, William
19. Rider, William
20. Cibber, Colley
21. Griffith, Mrs. (Elizabeth)
22. Fénelon, François de Salignac de…
23. Goldsmith, Oliver
24. Fenning, Daniel
25. Addison, Joseph
26. Walker, John
27. Voltaire
28. Garrick, David
29. Cibber, Theophilus
30. Enfield, William

Bibliography

Allan, D. (2010). *Commonplace Books and Reading in Georgian England*. Cambridge: Cambridge University Press.

Allen, T., Cooney, C., Douard, S., Horton, R., Morrissey, R., Olsen, M., Roe, G. and Voyer, R. (2010). Plundering Philosophers: Identifying Sources of the *Encyclopédie*. *Journal of the Association for History and Computing*,**13**(1).

Blair, A. (2011). *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven: Yale University Press.

Büchler, M., Crane, G., Mueller, M., Burns, P. and Heyer, G. (2011). One Step Closer To Paraphrase Detection On Historical Texts: About The Quality of Text Re-use Techniques and the Ability to Learn Paradigmatic Relations. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science.*

Büchler, M., Geßner, A., Berti, M. and Eckart, T. (2013). Measuring the Influence of a Work by Text Re-Use. In Dunn, S. and Mahony, S. (eds), *The Digital Classicist 2013*. London: University of London Institute of Advanced Studies, pp. 63-80.

Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C., Ossewaarde, R. and Jacobson, S. (2012). The Tesserae Project: Intertextual Analysis of Latin Poetry. *Literary and Linguistic Computing*, **28**(1).

Coffee, N., Gawley, J., Forstall, C., Scheirer, W., Corso, J., Johnson, D. and Parks, B. (2014). Modeling the Interpretation of Literary Allusion with Machine Learning Techniques. *Journal of Digital Humanities*, **3**(1).

Dacome, L. (2004). Noting the Mind: Commonplace Books and the Pursuit of the Self in Eighteenth-Century Britain. *Journal of the History of Ideas* **654**: 603-25.

Edelstein, D., Morrissey, R., and Roe, G. (2013). To Quote or not to Quote: Citation Strategies in the *Encyclopédie*. *Journal of the History of Ideas*, **74**(2): 213-36.

Horton, R., Olsen, M. and Roe, G. (2010). Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections. *Digital Studies / Le Champ numérique*, **2**(1).

Roe, G. (2012). Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research. In Meister, J.C., (ed), *Digital Humanities 2012*. Hamburg: University of Hamburg Press, pp. 345-47.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, **24**(4): 35-43.

Notes

[1] On our Digging into Data project, see http://diggingintodata.org/awards/2013, and on ECCO, see http://gdc.gale.com/products/eighteenth-century-collections-online/.

[2] See https://code.google.com/p/text-pair/.

[3] Gale's ECCO modules: History and Geography (17,950 works, reduced to 10,528); Social Sciences and Fine Arts (48,335 works, reduced to 30,498); Medicine and Sciences (15,636 works, reduced to 9,202); Literature and Language (53,351 works, reduced to 25,655); Religion and Philosophy (51,485 works, reduced to 29,962); Law (13,595 works, reduced to 7,726); and General Reference (5,198 works, reduced to 3,129).

[4] See http://www.textcreationpartnership.org/tcp-eebo/.

Implementation of a National Data Center for the Humanities (DaSCH)

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
University of Basel, Switzerland

Beat Immenhauser
beat.immenhauser@sagw.ch
Swiss Academy of Humanities and Social Sciences

Peter Fornaro
peter.fornaro@unibas.ch
University of Basel, Switzerland

Introduction

Up-to-date research in the humanities today depends as much on digital methods and digital data. However, the use of computer-based methods and online sources in the humanities still faces several challenges, including the difficulty of ensuring the longevity of research data, the lack of common basic services, inadequate standardisation of data formats, insufficient training in digital methods and best practices, and weak international Digital Humanities networks. Digital documents are accumulated, organised and annotated using electronic databases. However, the necessary infrastructure is most often established in a project- specific way and is not designed for the long-term preservation of data. After the completion of a research project, these digital resources quickly become unavailable if they, and the software and hardware they rely on, are not properly maintained. Keeping digital data accessible after the end of a project is costly in terms of money and labour and is usually not included in the project funding.

While the digitisation of analogue sources produces large numbers of digital documents, these documents usually have a simple structure. By contrast, the data produced during the research process is much more complex, consisting of interlinked information (databases, annotations etc.). Because of the complexity of this research data, it is very difficult to make it permanently available. However, there are several reasons for doing so:

Transparency:

As research data is the foundation on which published results are based, it be- comes necessary to have access to this data in order to evaluate the results.

Reuse:

New research projects can reuse existing research data to propose different answers to the same questions, or to ask entirely new questions, especially if the datasets from different projects can be linked.

Citability:

Digital sources may only be referenced in scientific texts if they can be accessed permanently without modification. The long-term accessibility of arbitrary digital objects (together with permanent links and unique object identifiers) is usually not possible.

Organisational form

The Swiss Academy of Humanites and Social Sciences (SAHSS) therefore decided to establish in collaboration with the Digital Humanities Lab (DHLab) of the University of Basel a new "national research infrastructure" (Data- and Service Center for the Humanities, DaSCH) which takes this kind of digital research data into custody and preserves the direct online access. The primary goals are:

- Long-term curation of research data
- Permanent access and reuse
- Services for researchers to support data life-cycle management

The secondary goals are:

- Promoting the digital networking of databases created in Switzerland or in other countries
- Carrying out a pilot project in close proximity to humanities research

- Collaboration and networking with other institutions on developing digital literacy

During a pilot phase lasting two years that ended in July 2015, the data of about 25 different research projects ranging from ancient history to musicology have been passed to new institution for preserving long term accessibility. In order to copy with such heterogeneous data, the platform has to be extremely flexible and versatile.

Since Switzerland is a highly federalist country, a balance between a central/decentral approach had to be chosen. We decided to form of a network that currently consists of several "satellite" nodes and a central office which acts as coordinator, main provider of technology and software development. The individual locations have a great deal of freedom to take local decisions (e.g. which research projects are considered important to be included in the platform). At each satellite location, it is necessary to have both a broad knowledge and experience available in humanities research as well as in IT and software development skills. The central office provides second-level support.

Technological issues

Our daily experience seems to suggest that digital data is quite volatile and unstable. Everybody who works with computers on any scale has suffered the unfortunate experience of data loss. In a recent interview, Vincent Cerf, often regarded as one of the "fathers of the internet", says he is worried that all the images and documents we have been saving on computers will eventually be lost: "Our life, our memories, our most cherished family photographs increasingly exist as bits of information – on our hard drives or in "the cloud". But as technology moves on, they risk being lost in the wake of an accelerating digital revolution." (Cerf, 2015) Thus, it appears that "long-term archival" and "digital" are diametrically opposed concepts. However, the digital domain offers some unique characteristics that allow the long-term preservation of digital data. However, guaranteeing long-term access to digital information remains a tedious and difficult process.

There are only a few fundamental methods for long-term preservation of digital data:

*Emulation*The software and to some extent the hardware of obsolete computer system can be emulated ("simulated") on modern computers. Thus data can be rendered using vintage software.

*«Eternal» media*The «eternal» media approach requires the digital data to be recorded onto the most robust and durable media available.

*Migration*In the context of long term archiving, migration is defined as the process of periodically copying digital data onto new, up-to-date storage media and, if required, converting the file formats to new, well-documented standard formats.

The OAIS reference model for a digital archive is based on the migration model. In addition to a formal process description, it also covers the ingest of data into the archive and the dissemination of archived data to a user. An important aspect of the OAIS reference model is the systematic approach to metadata that is distinguished between the metadata required to identify and find a «document», and the technical metadata required for the management of the migration processes. The OAIS approach can be adapted for complex «objects» such as relational databases or NoSQL-databases (e.g. using the SIARD-suite (Ohnesorge, 2015), a standard adopted by European PLANETS project and as Swiss eGovernment Standard eCH-0165), however in order to browse or use the data, the whole dataset has to be retrieved from the archive and converted back into a working RDBMS using the SIARD-Suite – a «quick overview» is not possible.

Complementary to the OAIS archival process model, *keep-alive archiving* keeps a system of data, data management and access methods online and permanently up-to-date. Whenever the technology evolves (e.g. a new stable version of the data management software or a new version of a file format is released), the whole system is migrated to conform to the new environment. The keep-alive archives are especially well suited to complex data such as databases which are accessed very frequently. However, there two fundamental problems with keep-alive archives:

If the data management system does not offer a method to record all changes, the history will be lost.

It is virtually impossible to keep each projects IT-infrastructure – especially the software – running forever. Each project uses its own software (Filemaker Version XY, MySQL, PHP, ruby, Excel, etc.) and data models. The adaption to the evolving technology would overwhelm each institution.

The DaSCH implements a modified keep-alive concept. It has chosen to use the Resource Description Framework (RDF, standardised by the W3C) as a common ground for representing the data. It provides a very simple but highly flexible representation of digital information. RDF allows the definition of ontologies which formalise the semantic relationship of digital objects. We defined a base ontology which implements some required basic concepts (e.g. time-stamp based versioning, annotations, access rights etc.). Starting from this base ontology, for each research project taken into custody a specific ontology is being derived. On delivery of the data, the original data structure is translated into this ontology preserving the important features and relationships of the data. This technological framework thus allows the «simulation» of almost any data models (relational databases, XML hierarchies, TEI-encoded texts, graph networks etc.) in a common infrastructure using open standards such as RDF, RDFS[1] and OWL[2].

The pilot phase has made it clear that project-specific access applications (such as online graph- ical user interfaces) have to be preserved. While this approach does not make it possible to directly reuse the original applications, it has been shown that is easy to re-implement their basic functionality as well as their look and feel.

Using the common platform, it is straightforward to create new tools and applications that reuse existing data by combining information from different datasets. Thus, new research methods can be implemented, e.g. using methods of «big data» analysis

Due to the success of the pilot phase where about 25 projects have been integrated, some with individual user interfaces, the Academy has decided to ask for funding. The request is awaiting the approval of the swiss national parliament.

Bibliography

Cerf, V. (2015). Interview on BBC http://www.bbc.com/news/science-environment-31450389 (accessed March 4th 2015).

Ohnesorge, K., Mérinat, T. and Büchler, M. (2015). SIARD Format Version 2.0, SFA | 2015-10-15 | DLM Forum 2015, Luxembourg, http://www.eark-project.com/resources/conference-presentations/dlm-oct15/37-siard2eark-1/file (accessed March 4th 2016).

Notes

[1] RDS-Schema for expressing simple ontologies.

[2] Web Ontology Language for expressing complex ontologies and relations.

Digicraft and 'Systemic' Thinking in Digital Humanities

Enrica Salvatori
enrica.salvatori@unipi.it
University of Pisa, Italy

In the AIUCD (*Associazione per l'Informatica Umanistica e la Cultura Digitale*) conference held in Bologna in September 2014 Manfred Thaller wondered "Are the Humanities an endangered or dominant species in the digital ecosystem?" (Thaller, 2014, also Thaller, 2012). The answer was not simple nor linear and directly involved the Digital Humanities (DH from now on) as a disciplinary and research field that was still poorly defined, DH hold the promise to bring out the Humanities from the Indian reserve where they are now confined, provided certain conditions will be met. In particular DH specialists should:

1. conceive of themselves as researchers and not as conversationalists;
2. strive for a vision;
3. change the epistemology of the Humanities;
4. drive technology and not be driven by it.

A few months after the conference Serge Noiret wrote on Digital History (one of my fields of investigation) trying to clarify what actually characterizes this subject within the wider field of DH, and - within the Digital History itself - what is the specific task of the Digital Public History (Noiret, 2015; see also Robertson, 2014). We could place Noiret's article completely under the first point of Thaller's list: Digital History and Digital Public History are clearly seen as areas of research and not merely as new forms of communication of old disciplines. Moreover he answered to items 2 and 3 too, essentially proposing a more accurate taxonomy of DH. In a way, he seems to answer Thaller's question with an accurate definition of some components of the meta-discipline itself.

I do not want to linger in this paper on the definition of DH as a whole nor of its components: so many authors in recent years have-debated to define what someone thinks to be (or could be) a discipline and others a research or work field (McCarty, 2005; Svensson, 2010). Every exercise of definition of a "new" area of research is, of course, useful, but at the same time it is potentially frustrating and risky. Frustrating because, as many authors and research centers have declared, despite its now long history, the DH is still an emerging field, and as well as an open, multifaceted, ever-changing one; risky, because each taxonomy of knowledge unavoidably builds walls and fences that encase the knowledge itself in a series of sterile boxes. This could be, in my modest opinion, the risk in Noiret's essay. It's more important to go beyond a possible but also difficult definition of DH and their several sub-disciplines, focusing our attention on items 2 and 3 of the Thaller list instead, namely on the need to have our own vision and on the importance to characterize DH in terms of the emerging changes of method in our daily research.

In particular I will try to connect the concepts expressed by both scholars, looking on the one hand to the recent history of DH in Italy (i.e. degree courses, associations, meetings) and on the other hand to my own research projects at the University of Pisa, especially inside the DH course degree (https://www.unipi.it/index.php/ects/ects?ects_id=IFU-L) and within the Digital Culture Laboratory (http://labcd.humnet.unipi.it/). Starting from some specific cases I wish to reason on the possible vision of the DH.

I will focus very briefly on some projects.

1. For what concerns *Epigraphical Studies, Public History and Education*:

- Epigrapisa: A re-reading partly driven and partly spontaneous of the epigraphic messages left over time in a city. Competence/knowledge at work: history, public history, epigraphy, paleography, writing, dramatize, processing images, audio and video, web design.
- Teaching (Digital) Epigraphy: a novel education

experience in teaching students to transcribe and interpret Roman inscribed lead tags, using a Digital Autoptic Process (DAP) in a Web environment. Competence/knowledge at work: history, education, epigraphy, paleography, writing, manipulating images, collaborative tools.
  - Pisa e l'Islam and Pisan Romanesque meets Contemporary America: two examples of historical web dissemination with a reasoning about both the potential and the limits of the medium to involve the audience. Competence/ knowledge at work: history, arts, archaeology, public history & archaeology, epigraphy, GIS, writing, dramatize, processing images audio and video, web design.
2. In the area of *Digital Public History*:
  - Tramonti. Itinerari tra generazioni lungo i crinali della Val di Vara a complex project aimed at enhancing the cultural heritage of an Italian rural valley through the active participation of residents. Competence/knowledge at work: history & archaeology, public history & archaeology, invented archives, education, writing, dramatize, GIS, manipulating images and videos, libraries, collaborative tools, web design, project management.
3. In the field of Digital Editions:
  - Codice Pelavicino Digitale: the digital edition of a medieval manuscript built to provide all services of the digital world and to invite the readers to actively participate. Competence/knowledge at work: history & public history, text encoding, philology, paleography, codicology, writing, manipulating images, collaborative tools, web design, project management.

By shortly describing these project I will not try to figure out what distinguishes them from each other, but, on the contrary, what characterizes all of them as Digital Culture projects and what they tell us about a possible vision of DH:

1. they are digital;
2. they are inevitably and necessarily interdisciplinary;
3. they are open;
4. they were built in a kind of new Renaissance workshop, a digital craft (DIGICRAFT).

- They are digital. This may seem trivial but it is not. These are projects "born digital" not because the digital world offers the most useful tools to achieve the same purpose in relation with the "real" world, but because they could not exist outside the incredible interaction between real and digital world that it is now our life. They are digital because they might not otherwise exist.
- Interdisciplinarity is compulsory. DH is a field unavoidably and profoundly interdisciplinary and we have to deal with each project as a complex set of activities and skills that crosses, by its true nature, several fields; this change of practice and approach implies by itself a methodological revolution, because it requires an organization of work similar to a Renaissance workshop (a DIGICRAFT), with an articulated division of labor in relation to several levels of skills, where education and training could be provided by the same learners, coordinated by a strong and mature central idea.
- Openness is a result and a choice. Working in a multidisciplinary team built upon research and with different tools, sustainability requires using open source tools, sharing data between individuals and giving everything to the public. Then Openness is a natural result, even it is also an ethical, political and philosophical choice as the Digital Manifesto 2.0 says: "the digital is the realm of the open, open source, open resources".
- A DIGICRAFT. In a Renaissance workshop it was possible to produce different objects: statues, paintings, goldsmith or less valuable coroplastic objects. Each handwork was a "project" that included on the one hand a strong artistic and cultural vision (meaning, style, function, purpose, style) and on the other hand a complex set of different techniques made by different workers with different levels of capacity. The owner of the shop (or the head-artist) had not necessarily to know each technique as an expert, but his employees could in many ways be superior and the various members of the workshop could learn from each other. The owner had to keep the team together with a well clear idea of the work itself.
- Likewise a DIGICRAFT could work (and actually does in our LabCD in Pisa) in the same way: each project is taken over as an interdisciplinary complex object that requires specific skills and different but profoundly related competences. The basis (first phases) of the work are often composed by students of the Bachelor and Master's degree in DH, who work, in the labCD, as interns or undergraduates. The work is directed by one "manager" but followed at various stages by experts, who assign specific tasks to the students, always ensuring an active connection among everyone in the team through the usual or more useful collaborative tools. While the work goes on, often happens that some student acquire, in a particular technique or phase, a greater capacity and knowledge than the others and then he/she becomes able to propose substantial changes in the work chain. The manager is not required - nor humanly could - to know every aspect in depth, nor to be fully aware of all the problems related to it, or to master each technique: however, he/she must be able:
  - to see always clearly the aim and the nature of the work;
  - to communicate effectively with everyone in the team.
- A "digicraft" is anywhere on a DH project teachers and students exchange knowledge and leverage this interaction to offer innovative and effective solutions, combining

the theoretical reasoning with practices and skills. This is possible only if the manager and the team share a common strong vision of what a DH project is, embracing a "systemic" or "organic" or "holistic" thinking of DH itself.

The core of DH is unitary and lies in the conviction that the digital turn has permeated every aspect of our lives as people and scholars modifying them deeply.

In the 70s of XXth century has increasingly gained ground a vision of Humanities Computing that kept almost unchanged the traditional disciplines within their rigid internal divisions and distinguished the humanist from the expert in information technology, hoping and promoting a dialogue between the two main areas (still in Fusi 2011, I, p. 1-2). Today this position is no longer sustainable. The web in first place and the web 2.0 in the second (but also the Big Data emerging field as well as the Data Visualization tools) have slowly but surely changed the research landscape especially demolishing the barrier between tools, methods and ways of sharing. We are obviously still in a transitional phase. Highly specialized sub-areas remain (and also in the future will exist) and obviously several scholars strive to better define the old / new digital disciplines (digital history, digital philology and so on), but there is also a complementary phenomenon pointing to an inclusive and unitary vision of DH.

From the perhaps limited but interesting Italian observatory I believe this change has affected both the terminology used in the establishment of centers, associations and degree programs (AIUCD, Digital Humanities degree, Digital Cultural Heritage, Arts and Humanities School), both the organisation of courses and meetings.

It 's more and more widespread the awareness that we are a new type of scholar (and graduate, and PhD), the digital humanist, someone who has a mixed formation, an open mind, is able to master both languages and the main methodological issues of the two areas without considering one serving the other.

In doing so, we need to maintain the epistemological strictness that each discipline involved in the DH has developed over time: commingling does not mean carelessness or inaccuracy; but in the same time we have to claim the change or the changes in each methodology in order to build a new global epistemology.

In order to do this the digital humanist has to embrace a "systemic" or "organic" or "holistic" thinking of the humanities, leave the enclosure of the academic fields and get away from the temptation to create an old-new rigid taxonomy.

Speaking of "systemic" or "organic" or "holistic" thinking / view (I'm sorry for the repetition but the adjectives organic and holistic mean different things not only in different languages but mainly in different subjects), I refer to the epistemological approach that has emerged in some areas of the research over the past thirty years and which tends to oppose the reductionist approach flourished since the seventeenth century onwards and imposed in almost all sectors of the so-called "hard sciences" (Capra-Luisi 2014). As we know reductionism believes that studying in depth a peculiarity of a phenomenon and understanding it completely it will be possible, by progressive addition of discoveries, illuminate the entire system. The reductionist approach has been, as we know, the basis for the scientific revolution of the modern age, but it also led in the eighteenth and nineteenth century to an exasperate

fragmentation of the fields of scientific research. This phenomenon has also heavily influenced the Humanities,often creating absurd barriers and hyperspecialized languages, that have closed researches in several walled gardens. I believe that this long wave has exhausted its strength and that precisely the DH can reverse the trend. Now a new methodological approach have arisen alongside the reductionist thinking, considering the "system", the "whole", something more and different than the sum of its components. The "systemic thinking" reasons in terms of relationships, networks, patterns of organizations and processes; it proposes a change of paradigms: from the vision of the world as a machine to the world as a network; it takes account of the fundamental interdependence of all phenomena.

This change of paradigms could and should affect the DH as well for the reasons listed above, promoting a systemic view of this meta-discipline and therefore pushing Digital Humanists to deeply transform the old practice of work.

Bibliography

AIUCD Associazione per l'Informatica Umanistica e la Cultura Digitale, http://www.umanisticadigitale.it/ (accessed 27 February 2016).

Boonstra O., Breure L., Doorn P. (2006). *Past, Present and Future of Historical Information Science.* Amsterdam.

Capra F., Luisi P.L. (2014). *The Systems View of Life: A Unifying Vision,* Cambridge; italian edition *Vita e natura. Una visione sistemica.* Sansepolcro.

The Digital Humanities Manifesto 2.0 http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf

Fusi D. (2011). *Informatica per le scienze umane,* I, Elementi, Roma.

McCarty W. (2005). *Humanities Computing,* London-New York.

Noiret S. (2015). Digital Public History: bringing the public back. *Public History Weekly*, 3:13 http://public-history-weekly.oldenbourg-verlag.de/3-2015-13/digital-public-history-bringing-the-public-back-in/

Robertson S. (2014). The Differences between Digital History and Digital Humanities. *CHNM Blog Post* (23/1/2014), http://drstephenrobertson.com/blog-post/the-differences-between-digital-history-and-digital-humanities/ (accessed 27 February 2016)

Salvatori E. (2015). Il patrimonio genetico della storia digitale (e le nostre paure), *Appunti di viaggio tra il medioevo e la* *cultura digitale* (05/03/2015), http://esalvatori.hypotheses.org/211 (accessed 27 February 2016).

Salvatori E. (2015). L'identità dell'Informatico Umanista e la visione sistemica , *Appunti di viaggio tra il medioevo e la cultura digitale* (23/01/2015) http://esalvatori.hypotheses.org/204 (accessed 27 February 2016).

Svensson P. (2010). The Landscape of Digital Humanities, *Digital Humanities Quarterly*, **4**(1).

Thaller M. (2012). *Controversies around the Digital Humanities: An Agenda, Historical Social Research / Historische Sozialforschung*, **37**(3): 7-23, http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html

Thaller M. (2014). *Keynote address*. AICUD 2014, http://aiucd2014.unibo.it/ (accessed 27 February 2016).

Using Computer Numerical Control Techniques to Prototype Media History

Jentery Sayers
jentery@uvic.ca
University of Victoria, Canada

Tiffany Chan
tjychan@uvic.ca
University of Victoria, Canada

New technologies can now be used to fabricate old ones. With rapid prototyping techniques, a nineteenth-century mechanism can be downloaded from an online library, translated into code, fed to a 3-D printer, and used to repair a watch, all in about an hour. While such 3-D fabrication techniques tend to fetishize objects, this talk proposes an alternative: "prototyping the past," or remaking technologies that no longer function, no longer exist, or may have only existed as fictions, illustrations, or one-offs. Conceptually, prototyping the past explains why technologies matter by approaching them as representations *and* agents of history. Practically, it creates media that function simultaneously as evidence *and* arguments for interpreting the past. Yet most important, it does more than re-contextualize media history in the present. It integrates that history into the trajectories of design practice.

This full circuit through digital and analog manufacturing corresponds rather provocatively with a "materialist turn" in media studies, where, since at least the late 1990s, scholars have stressed the importance of media as physical entities that resist or diffract interpretations as often as they facilitate them.[1] Writing about phonography, Lisa Gitelman asserts, "At certain levels, media are very influential, and their material properties do (literally and figuratively) *matter*, determining some of the local conditions of communications amid the broader circulations that at once express and constitute social relations" (2006, p. 10). Somewhere between technological determinism[2] and social constructivism,[3] this assertion accounts for how—as both representations *and* agents of history—media play an active role in reproducing the past. Following Gitelman, in the following paragraphs I situate media history in rapid prototyping research, looking specifically at the effects of fabricating tactile models to "prototype the past." While instrumental uses of rapid prototyping tend to fetishize objects, encourage commodification, foster nostalgia, or ignore the political dimensions of material culture, I argue that rapid prototyping can help researchers demonstrate how social and cultural complexity is intricately entangled with the historical particulars of design.[4] Ultimately, what distinguishes rapid prototyping from most other approaches to media history is its investment in not only translating between 2-D and 3-D modes of trial-and-error production but also communicating across a spectrum of media that operate simultaneously as evidence and arguments for future work. Put simply, prototyping the past is more than re-contextualizing media history in the present. It is an opportunity to meaningfully integrate that history into the social, cultural, and ethical trajectories of design practice.

More common in engineering and architecture than the humanities, rapid prototyping entails producing materials through a combination of computer numerical control (CNC) machines—such as 3-D printers (additive manufacturing) and routers (subtractive manufacturing)—with manual approaches to wood, paper, clay, etc. The aim is to subject a 3-D model to repeated feedback throughout the development process. In this sense, the design cycles are small, not grand. Also, the models are versioned. Instead of working toward a single model, multiple models are maintained throughout production. This approach is steeped in "design-in-use," which privileges situated activity over some ideal model or user (Botero, 2013). Through design-in-use, a prototype is treated like a congealed dialogue between interested groups.

With the above in mind, this talk explains why scholars of media history may wish to prototype the past. It builds on especiwork by Kraus (2009), Balsamo (2011), Ratto (2011), Buechley (2012), Perner-Wilson (2012), Ames (2014), Hjorth (2014), Jungnickel (2014), and Rosner (2014)., and It also they in part corresponds with arguments published in "New Old Things" (2012), by Elliott, MacDougall, and Turkel. There, Elliott *et al.* argue that "matter [is] a new medium for historical research. Working with actual, physical stuff offers the historian new opportunities to explore the interactions of people and things" (2012, p. 122). Also, rapid prototyping may frame prototypes may be understood as *situations* for interpretation, without creating exact reproductions of historical artefacts (127). By extension, using matter as a medium for historical research need notneed not fetishize the past. Instead, itt becomes a time and space to interpret the material intricacies of technological design, both now and then.

Perhaps most important, prototyping the past allows scholars to remake technologies that no longer function, no longer exist, or may have only existed as illustrations, fictions, patents, or—fittingly enough—prototypes. TheseOne appeal of prototyping the past is that remade technologiesremade technologies may be circulated as tangible reminders of what was forgotten or destroyed. Yet prototyping the past also affords critiques of what is ready to hand. That is, *it refuses to take historical materials at face value*. Through trial and error, it tests the plausibility of historical claims. After all, what is depicted in a text may contain redactions, deliberate omissions, purposeful obfuscations, or accidental occlusions. Using historical materials to prototype a technology amplifies the meaningfulness of these absences.

Put this way, prototyping the past is intertwined with close reading. However, its emphasis on physically remaking historical technologies expands hermeneutics to include the centrality of translation and tacit knowledge toin media history. By re-contextualizing historical technologies in the present, prototyping also accentuates differences across time. What was once an innovation in the 1860s becomes a relic in 2015. Alternatively, these differences across time may turn things of the past into the stuff of present-day current curiosity., thus complicating distinctions between old and new.

Rather than transcending such differences, prototyping the past grounds media history in a particular thing and the interpretations it affords. Following the work of Karen Barad, such grounding posits prototypes as entanglements of meaning with matter by attending to the substance of "fine-grained details" (Barad , 2007, p. 90). Here, neither meaning nor matter can be relegated to a concept or abstraction. Again, situations are significant. And prototyping reminds scholars of that significance. It is an embodied process involving frustration and surprise, trial, and error, and it highlights how technologies do not emerge effortlessly from the brilliant minds of inventors or makers. Prototyping also reminds scholars that 1) the sources of meaning are forever unstable and under dispute, 2) historical materials are not complete"total" works but rather compositions of parts that change—degrade, rot, morph, warp, break, or swell—over time, 3) prototypes can be de- and remanufactured without significant material consequence, and 36), as noted earlier, materials resist or diffuse as many interpretations as they facilitate.

Example Application: Kits for Cultural History

Based at the Maker Lab in the Humanities ("Lab") at the University of Victoria, the Kits for Cultural History ("Kits") remake technologies from the past, package them in bespoke containers, contextualize them with historical materials, and encourage people to experiment with them. Comparable to Heathkits, the Kits include components and guides for assembly. However, the guides are steeped in cultural history and do not assume a single approach to assembly. By design, this resistance to uniformity is essential, since the Kits focus on technologies that are inaccessible today. These technologies are not found in galleries, libraries, archives, or museums; they no longer function as they once did; or they were never actually built or mass-manufactured. Such inaccessibility necessarily entails a degree of uncertainty where research is concerned. Rather than approaching this uncertainty at a remove, *the Kits prototype absences in the historical record and prompt audiences to examine the contingencies of that record*. Anchored in design-in-use, this method approach presents prototypes as negotiations or situations for interpretation, not replicas.

The Kits' design cycle may be visualized as follows:

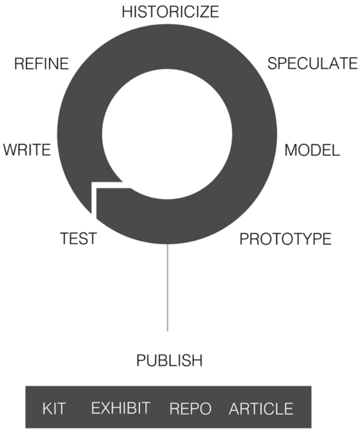


Figure 1: Design Cycle for the Kits

Once the Lab selects a technology for remaking, we historicize it through archival materials. Informed by existing theoriesy, we then speculate about absences in the record and determine how those absences might manifest in tactile form. Next, we model, fabricate, and assemble the technology's component parts into prototypes, which we test and share with other researchers. After feedback, the Lab writes about the process and related history. When bundled together, the writing, prototyping, and testing refine ourthe research, and the cycle is repeated until we deem a Kit persuasive. Once a Kit is ready for circulation, we publish it in tactile form, as a repository of digital files, and through an exhibit. With these, the Lab also authors articles about the Kit's contribution to media history, theorhistory, and design. We treat these publications—the tactile Kit, repository, exhibit, and article—equally as elements of scholarly communication.

Throughout this cycle, we ask several questions of what we are prototyping: 1) Who made it? For whom? When? 2) How was it made? How did it work? How was it used? 3) Do any instances of it exist? If so, where are they? Can they be handled? 4) Under what assumptions was it made and used? 5) How might prototyping it shape design in the future?

We have used this process to produce three Kits thus far: an Early Wearable Kit (for an electro-mobile skull stick-pin from 1867), an Early Magnetic Recording Kit (for experiments involving steel wire, telephones, and carriages during the late 1890s), and an Early Optophonics Kits (for a reading aid patented in 1919).

Findings: Rapid Prototyping and Media History, Together

The Kits support the following observations about the articulation of CNC techniques with media history: 1) prototyping the past demands methods from the humanities, engineering, and fine arts; 2) prototyping is not always futurist or restricted to forecasting; 3) 3-D media such as tactile models are not more persuasive than 2-D media such as illustrations; both include exaggeration and omission; 4) history remains inaccessible even with access to physical materials, which neither resolve issues of absence nor guarantee certainty about the past; 5) prototyping the past may actually resist nostalgia; 6) as with any research method, prototyping is not immediate and cannot access "real history"; 7) prototyping may be premised on *not* replicating history—on what we should *not* repeat; 8) prototyping tests suspicions of history by grounding them in fine-grained details of matter and meaning; and 9) prototyping the past need not aim for a rational history without remainders. Instead, it can recognize how the technologies we use to reproduce history exceed our control and understanding. Indeed, the speculative elements of prototyping can be anchored in the specificities surrounding historical absences—of what we cannot prove or do not know for sure but certainly shapes us.

Acknowledgments

The Social Sciences and Humanities Research Council, the Canada Foundation for Innovation, and the British Columbia Knowledge Development Fund have supported this research.

Bibliography

Balsamo, A. (2011). *Designing Culture*. Duke UP.

Barad, K. (2007). *Meeting the Universe Halfway*. Duke UP.

Botero, A. (2013). *Expanding Design Space(s).* Aalto U.

Buechley, L. and Perner-Wilson, H. (2012). Crafting Technology. *ACM Trans. Comput.-Hum. Interact.*, **19**(3).

Elliott, D., MacDougall, R. and Turkel, W. J. (2012). New Old Things. *Canadian Journal of Communication*, **37**(1).

Jungnickel, K. and Hjorth, L. (2014). Methodological Entanglements in the Field . *Visual Studies*, **29**(2).

Kraus, K. (2009). Conjectural Criticism. *Digital Humanities Quarterly*, **3**(4).

Pinch, T. J. and Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, **14**(3): 399–441.

Postman, N. (1992). *Technopoly: The surrender of culture to technology* (Reprint edition). New York: Vintage.

Ratto, M. (2011). Critical Making. *The Information Society*, **27**(4).

Rosner, D. K. and Ames, M. (2014). Designing for Repair? *17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing.*

Notes

[1] For examples of this work, see Kittler (1999), Gitelman (1999, 2006, 2014), Bowker and Star (2000), Sterne (2003, 2012), Galloway (2006), Kirschenbaum (2008), Vismann (2008), Chun (2011), Ernst (2012), Parikka (2013, 2015), and Starosielski (2015).

[2] Technological determinism asserts that technologies cause cultural and social phenomena. For more, see Postman (1992).

[3] Social constructivism asserts that cultural and social phenomena precede or shape the development of technologies. For more on social constructivism, see Pinch and Bijker (1984). For an overview of technology and cultural criticism, see Sayers (2014).

[4] Put this way, the combination of media history with rapid prototyping corresponds with how Wendy Chun describes the aims of the *New Media, Old Media* anthology, which focuses on the "actualities of the media itself" and "the experience ... of being entangled within it" (2005, p. 9).

Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels

Christof Schöch
christof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Daniel Schlör
daniel.schloer@informatik.uni-wuerzburg.de
University of Würzburg, Germany

Stefanie Popp
stefanie.popp@uni-wuerzburg.de
University of Würzburg, Germany

Annelen Brunner
brunner@ids-mannheim.de
Institut für deutsche Sprache, Mannheim

Ulrike Henny
ulrike.henny@uni-wuerzburg.de
University of Würzburg, Germany

José Calvo Tello
jose.calvo@uni-wuerzburg.de
University of Würzburg, Germany

Introduction

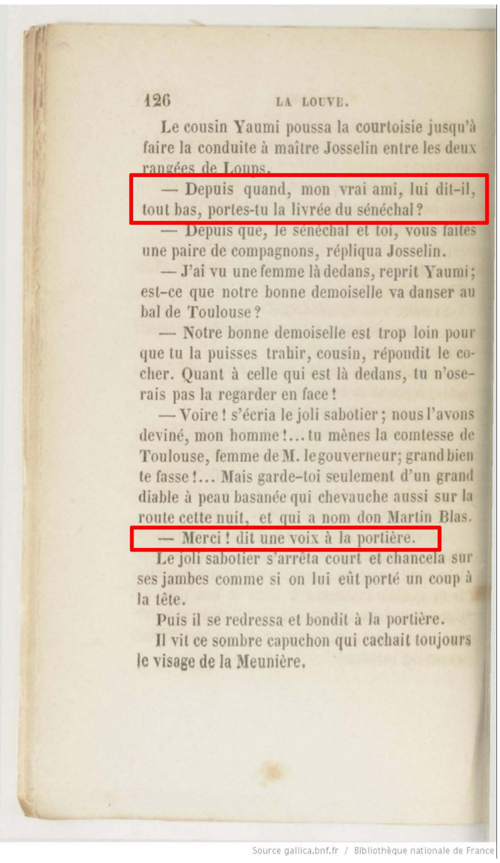
In fictional prose narrative such as novels and short stories, various forms of speech, thought, and writing representation are ubiquitous and have been studied in great detail in linguistics and literary studies. However, beyond quotation marks, what are linguistic markers of direct speech? And just how ubiquitous is direct speech really? Is there systematic variation in the amount of direct speech over time or across genres? Especially for the field of French literary history, where typography is not a reliable guide, we really don't know.

This is regrettable, because being able to quickly and automatically detect direct speech in large collections of literary narrative texts is highly desireable for many areas in literary studies. In the history of literary genres, this allows to observe distributions and evolutions of a fundamental, formal aspect of the novel on a large scale. In narratology, differentiating narrator from character speech is a precondition for more detailed analyses of narrator speech, e.g. with regard to text type (descriptive, narrative, argumentative text). And in authorship attribution, it hereby becomes possible to discard character speech from a set of novels and perform authorship attribution on the narrator speech only, something which may improve attribution.

Against this background, the work presented here addresses both the question of how to identify direct speech in French prose fiction and that of how prevalent direct speech is in different subgenres of the nineteenth-century French novel.

Aims and hypotheses

Our first aim has been to use machine learning to automatically identify direct character speech in a small collection of French-language fictional prose. This is less trivial than it seems to be since in the French typographical tradition, direct speech is usually not marked with opening and closing quotation marks (figure 1). Rather, a long hyphen usually indicates the beginning of direct speech, whereas the end is left unmarked. In figure 1, the first highlighted direct speech continues after the insertion revealing who has just spoken ("lui dit-il, tout bas,"; *he quietly said to him*). In the second example, the direct speech ends after the speaker has been indicated ("dit une voix à la portière"; *said a voice at the door*). Our hypothesis is that there are enough linguistic markers of direct speech to make it possible to identify it automatically and reliably (for an overview of such markers, see Durrer, 1994).



126 LA LOUVE.

Le cousin Yaumi poussa la courtoisie jusqu'à faire la conduite à maître Josselin entre les deux rangées de Loups.

— Depuis quand, mon vrai ami, lui dit-il, tout bas, portes-tu la livrée du sénéchal ?

— Depuis que, le sénéchal et toi, vous faites une paire de compagnons, répliqua Josselin.

— J'ai vu une femme là dedans, reprit Yaumi ; est-ce que notre bonne demoiselle va danser au bal de Toulouse ?

— Notre bonne demoiselle est trop loin pour que tu la puisses trahir, cousin, répondit le cocher. Quant à celle qui est là dedans, tu n'oserais pas la regarder en face !

— Voire ! s'écria le joli sabotier ; nous l'avons deviné, mon homme !... tu mènes la comtesse de Toulouse, femme de M. le gouverneur ; grand bien te fasse !... Mais garde-toi seulement d'un grand diable à peau basanée qui chevauche aussi sur la route cette nuit, et qui a nom don Martin Blas.

— Merci ! dit une voix à la portière.

Le joli sabotier s'arrêta court et chancela sur ses jambes comme si on lui eût porté un coup à la tête.

Puis il se redressa et bondit à la portière.

Il vit ce sombre capuchon qui cachait toujours le visage de la Meunière.

Source gallica.bnf.fr / Bibliothèque nationale de France

Figure 1: Detail from Paul Féval, La Louve, 1857, p. 126 (Source: http://gallica.bnf.fr/ark:/12148/bpt6k6366934b)

Our second aim has been to use the best-performing algorithm to identify direct speech in a larger collection of French nineteenth-century novels and to study its distribution. Here, we hope to detect significant differences in the proportion of direct speech found in different novelistic subgenres. Research about other literary traditions supports this hypothesis (e.g. Allison et al., 2011).

State of the Art

Speech, thought and writing representation are common topics in narratology and stylistics (Genette, 2008, Leech/Short, 2007). Semino and Short's 2004 quantitative study finds that direct representation is clearly the most frequent type in their English fiction sub-corpus. Brunner 2015 confirms this trend for her corpus of German short narratives. Here, the percentage of sentences containing direct speech is about 35% and varies widely over different texts (2%-72%).

Frequently, speech representation recognition is an auxiliary step to other tasks, e.g. knowledge extraction or speaker recognition (Krestel at al., 2008, Elson and McKeown, 2010, Iosif and Mishra, 2014, Sarmento/Nunes, 2009). Weiser and Watrin 2012 used a rule-based approach to extract unmarked quotations in French newspaper texts with success rates of 0.745-0.789. Brunner 2015 focuses on speech, thought and writing representation in German short literary narratives. Using machine learning with random forests, she reports an F1 score of 0.87 for direct speech in a sentence-based cross-validation.

Data

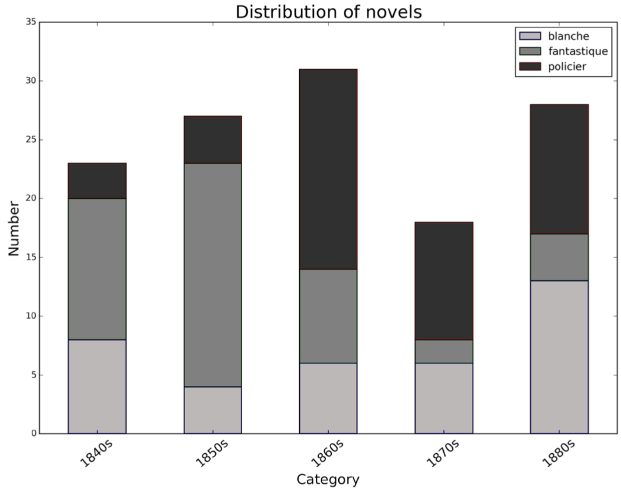


Figure 2: Distribution of novels per subgenre and decade

Our text collection contains 127 French novels published between 1840 and 1889. Three generic subsets can be distinguished, each of which is represented by approximately 40 texts: general novelistic fiction (so-called 'littérature blanche') is contrasted with specific subgenres, crime fiction ('policier') and fantastic novels (see figure 2). The narrative perspective is largely heterodiegetic.

Methods

Manual Annotation

To obtain a gold standard, 40 chapters from 20 different novels were randomly chosen from the collection and annotated manually. 5734 sentences were marked as either containing direct speech or not containing any direct speech; the former also include mixed sentences.

Preprocessing

To prepare feature generation, preprocessing was performed, the pipeline consisting of the Stanford CoreNLP-Tokenizer and Sentence-Splitter, as well as the TreeTagger for POS-Tagging and Lemmatization.

Feature generation

We modeled 81 features which we believed to be useful cues for the classification task (see the annex for a ranked list). Features are generated on a sentence-based level and can be divided into different categories:

- Character-based: e.g. long hyphen marks, exclamation marks, question marks.
- Lexical: e.g. deictic expressions, interjections.
- Semantic: categories of verbs, from WordNet and the French equivalent WOLF: e.g. verbs of motion or perception.
- Morphological: e.g. part-of-speech, verb-tense, lemma.
- Syntactic features: e.g. number of commas, sentence length.

Classification

For the binary classification task (sentences containing vs. not containing direct speech), we used an annotation and classification framework developed by Markus Krug (Würzburg) wrapping LibSVM Support-Vector-Machine (Chang and Lin, 2011), Maximum Entropy (Nigam et al., 1999) and Naïve Bayes (John and Langley, 1995) and implemented in MALLET (McCallum 2002). Random Forest (Breiman, 2001) and JRip (Cohen, 1995) were applied using Weka. All experiments were validated using 10-fold cross-validation unless otherwise stated.

Error analysis

The machine learning algorithms' incorrect assignments on the gold standard (false positives and false negatives) were manually analyzed in order to detect the errors' underlying causes.

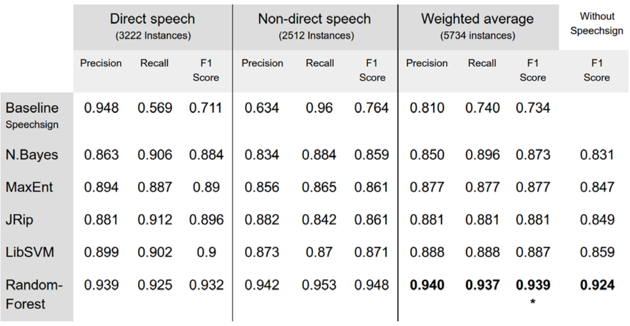
Automatic tagging of unseen texts

Using the best-performing model, all sentences in the text collection were tagged for containing direct speech or not. The distribution of ratios of direct speech / non-direct speech was calculated for the three subgenres and five decades covered by the collection. Performance on these unseen texts was checked manually on a random sample. (We sampled 2300 sentences, i.e. 100 random sentences each from a sample of 23 novels stratified by ratio of direct speech.)

Results and Discussion

Recognition of direct speech

Table 1 depicts the performance for different conditions. Our baseline is using the speech sign (i.e. the long hyphen) as the only feature, which yields an F1 score of 0.734. Random-Forest performs best, with an F1 score of 0.939, which we consider to be an impressive result. Even when excluding the speech sign from the features, we still reach an F1 score of 0.924, much better than the hyphen alone.



	Direct speech (3222 Instances)			Non-direct speech (2512 Instances)			Weighted average (5734 instances)			Without Speechsign
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	F1 Score
Baseline Speechsign	0.948	0.569	0.711	0.634	0.96	0.764	0.810	0.740	0.734	
N.Bayes	0.863	0.906	0.884	0.834	0.884	0.859	0.850	0.896	0.873	0.831
MaxEnt	0.894	0.887	0.89	0.856	0.865	0.861	0.877	0.877	0.877	0.847
JRip	0.881	0.912	0.896	0.882	0.842	0.861	0.881	0.881	0.881	0.849
LibSVM	0.899	0.902	0.9	0.873	0.87	0.871	0.888	0.888	0.887	0.859
Random-Forest	0.939	0.925	0.932	0.942	0.953	0.948	**0.940**	**0.937**	**0.939** *	**0.924**

Table 1: Performance (10-fold cross-validation on the gold standard)

After inspecting the models, it becomes clear that only very few features carry strong cues for direct speech, namely (and unsurprisingly) the initial long hyphen. Most other features, taken separately, carry weak signals in either direction, but become relevant in combination.

Error analysis reveals that incorrect assignments (false positives and negatives) are frequently due to imperfect sentence segmentation. Several features which have been previously used to define and recognize direct speech (question / exclamation marks, interjections, verbal tenses) also cause incorrect assignments, especially in the context of homodiegetic narration, where the narrator is somewhat involved in the plot so that his narrator speech is similar to direct speech. Finally, letters are sometimes mistaken for direct speech, which makes sense given that in most of them, one person addresses one or several other people.

Distribution of direct speech in the corpus

We applied the best-performing algorithm (Random Forest) to the entire text collection. Evaluation shows a certain drop in performance, with a weighted average success rate of 0.844, indicating less-than-perfect generalization. We noted a welcome absence of any strong bias for either direct or non-direct speech. Our results suggest that the average proportion of direct to non-direct speech across the collection is 61% sentences with direct speech (and 39% without direct speech).

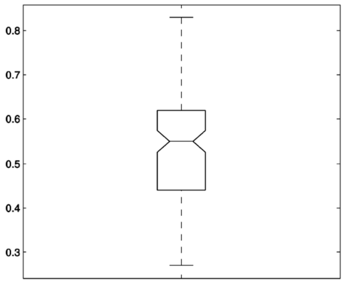


Figure 3: Ratio of direct to non-direct speech in 127 novels

While variance is considerable (see figure 3), the proportion of direct speech in French nineteenth-century novels is overall much higher than expected (and higher, for example, than the 35% reported by Brunner 2015 for German novellas).

Figure 4 shows that both fantastic novels and crime fiction have a significantly higher median for proportion of direct-speech than 'littérature blanche', but do not differ significantly from each other (for significance tests, we used the non-parametric Kruskal-Wallis test at a significance level of 1%).

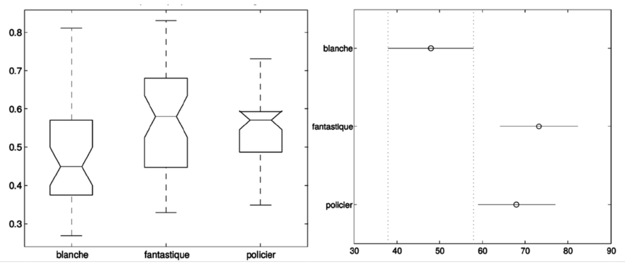


Figure 4: Distribution (left) and significance (right) of direct to non-direct speech ratios across three subgenres

Figure 5 shows that only the ratios for the 1850s and the 1880s have a significantly differing level. However, because the decades do not have perfectly balanced subgenre proportions, this is probably due to a subgenre imbalance rather than an effect of the time period.

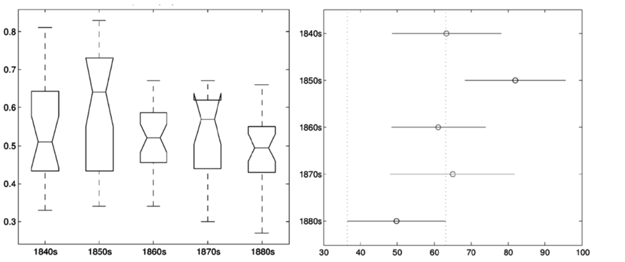


Figure 5: Distribution (left) and significance (right) of direct to non-direct speech ratios across five decades

Conclusions and Future Work

Using a wide range of linguistic markers allows the reliable identification of direct speech, even in the absence of clear typographic markers. Performance is excellent to good (F1-score of 0.94 on the gold standard, weighted average success rate of 0.844 on unseen texts). Using our method reveals that nineteenth-century French novels contain a large proportion of sentences with direct speech (61% on average). Also, there are previously unseen differences in direct speech proportion for subgenre, but not for time period.

For future work, we plan to use several strategies to improve performance. One is to add more sequential information to our set of features. Examples include the position, inside a sentence, of certain lexical or typographical features as well as linguistic cues preceding and following direct speech. Also, we plan to expand our corpus to make it more balanced in terms of genres and decades. This will allow us to discover genre-related patterns of interest to literary historians in a more reliable manner and assess their significance with more confidence.

Supplementary material

Supplementary material can be found at: https://github.com/cligs/projects/tree/master/2016/dh.

Annex A: Features used

List of features used, sorted by descending rank by a one-rule classifier.

```
average merit average rank attribute
74.028 +- 0.168 1 +- 0 79 SPEECHSIGN
71.743 +- 0.16 2 +- 0 57 VER:impf
65.847 +- 0.234 3 +- 0 54 VER:pres
63.893 +- 0.155 4 +- 0 55 VER:simp
63.248 +- 0.136 5 +- 0 6 PUNCMARKDOT
59.48 +- 0.12 6 +- 0 29 MATCHINGPPER_SON
58.835 +- 0.094 7.7 +- 0.64 30 MATCHINGPPER_SES
58.695 +- 0.208 8.1 +- 0.94 24 MATCHINGPPER_IL
58.713 +- 0.104 8.4 +- 0.92 35 VERB_MOTION
58.364 +- 0.083 10.6 +- 0.49 28 MATCHINGPPER_SA
58.344 +- 0.417 10.8 +- 1.78 7 SENTENCELENGTH
58.172 +- 0.078 11.7 +- 0.46 61 VER:subi
57.492 +- 0.091 14 +- 1.41 25 MATCHINGPPER_ELLE
57.422 +- 0.103 14.5 +- 1.36 44 VERB_PERCEPTION
57.387 +- 0.248 14.9 +- 1.51 50 INNERSUBCLAUSE
57.356 +- 0.4 15.8 +- 2.09 48 UNKNOWNLEMMA
57.213 +- 0.07 16.5 +- 1.02 31 MATCHINGPPER_LEUR
57.143 +- 0.162 17.3 +- 1.1 60 VER:ppre
56.672 +- 0.042 20.2 +- 0.98 36 VERB_BODY
56.672 +- 0.115 21 +- 1.84 52 VER:cond
56.62 +- 0.136 21.7 +- 2.1 40 VERB_EMOTION
56.567 +- 0.072 22.3 +- 1.19 26 MATCHINGPPER_ILS
56.497 +- 0.033 23.9 +- 1.3 41 VERB_COGNITION
56.428 +- 0.044 25 +- 1 46 VERB_CONSUMPTION
56.201 +- 0.005 34.5 +- 4.06 20 MATCHINGPPER_VOTRE
56.339 +- 0.176 35.4 +-18.69 32 COMMAS
56.201 +- 0.005 35.8 +- 4.19 21 MATCHINGPPER_VOS
56.201 +- 0.005 35.8 +- 6.4 22 MATCHINGPPER_TOI
56.201 +- 0.005 36.3 +- 4.2 17 MATCHINGPPER_TES
56.201 +- 0.005 37.6 +- 7.35 5 PUNCMARKCOLON
56.195 +- 0.018 37.7 +-13.33 18 MATCHINGPPER_NOTRE
56.201 +- 0.005 38.2 +- 3.16 23 MATCHINGPPER_MOI
56.424 +- 0.296 38.4 +-25.85 47 VERB_COMMUNICATION
56.201 +- 0.005 38.6 +- 6.45 4 PUNCMARKEXCL
56.201 +- 0.005 38.7 +- 3.44 16 MATCHINGPPER_TON
56.201 +- 0.005 39.4 +- 4.82 15 MATCHINGPPER_TA
56.201 +- 0.005 39.6 +- 6.45 3 PUNCMARKQUSTION
56.201 +- 0.005 40.2 +- 8.81 8 MATCHINGPPER_JE
56.201 +- 0.005 41.8 +-10.17 9 MATCHINGPPER_TU
56.201 +- 0.005 43.5 +- 9.19 10 MATCHINGPPER_NOUS
56.201 +- 0.005 43.5 +- 2.84 13 MATCHINGPPER_MON
56.201 +- 0.005 44.6 +- 4.43 12 MATCHINGPPER_MA
56.201 +- 0.005 44.7 +- 6.47 11 MATCHINGPPER_VOUS
56.261 +- 0.436 45.6 +-27.28 1 AmmountOfPPER
56.201 +- 0.005 45.8 +- 9.65 75 INTERJECTION_FI
56.201 +- 0.005 48 +-14.72 76 INTERJECTION_HEP
56.201 +- 0.005 50.2 +- 9.34 73 INTERJECTION_EH
56.201 +- 0.005 50.2 +- 6.27 74 INTERJECTION_EUH
56.201 +- 0.005 51.3 +- 3.66 81 INTERJECTION_MADAME
56.203 +- 0.08 51.3 +-23.56 37 VERB_COMPETITION
56.201 +- 0.005 52.1 +-15.75 58 VER:infi
56.201 +- 0.005 52.3 +-16.54 56 VER:futu
56.201 +- 0.005 53 +- 8.91 78 INTERJECTION_OUSTE
56.201 +- 0.005 54.7 +-17.43 34 VERB_CONTACT
56.162 +- 0.116 55.6 +-18.7 33 VERB_WEATHER
56.201 +- 0.005 56.9 +- 6.55 64 INTERJECTION_OH
56.201 +- 0.005 57.3 +- 4.5 63 INTERJECTION_AH
56.135 +- 0.087 58 +-24.31 19 MATCHINGPPER_NOS
56.193 +- 0.015 58.7 +-13.46 77 INTERJECTION_OUF
56.201 +- 0.005 59.3 +- 9.42 67 INTERJECTION_HÉLAS
56.143 +- 0.07 59.6 +-16.69 14 MATCHINGPPER_MES
56.201 +- 0.005 59.9 +- 4.5 42 VERB_STATIVE
56.201 +- 0.005 60.2 +- 2.64 62 VER:subp
56.201 +- 0.005 60.6 +- 8.39 71 INTERJECTION_CHUT
56.201 +- 0.005 62 +- 6.36 70 INTERJECTION_HEM
56.193 +- 0.015 62.5 +-10.87 66 INTERJECTION_HEIN
56.197 +- 0.011 62.6 +- 8 65 INTERJECTION_HÉ
56.201 +- 0.005 62.7 +- 5.87 51 DEIKTIKA
56.201 +- 0.005 63 +- 4.07 80 INTERJECTION_MONSIEUR
56.201 +- 0.005 63 +-11.79 53 VER:impe
56.005 +- 0.298 63.8 +-24.78 38 VERB_POSSESSION
56.201 +- 0.005 64 +- 5.67 39 VERB_SOCIAL
56.201 +- 0.005 64.3 +- 4.86 45 VERB_CHANGE
56.197 +- 0.013 64.7 +- 8.74 68 INTERJECTION_BAH
56.201 +- 0.005 64.7 +- 5.27 59 VER:pper
56.197 +- 0.015 65.2 +- 7.08 69 INTERJECTION_HOLÀ
56.139 +- 0.062 66.7 +-20.16 27 MATCHINGPPER_ELLES
56.183 +- 0.008 71.8 +- 5.23 72 INTERJECTION_BRAVO
56.005 +- 0.121 74.2 +- 9.41 43 VERB_CREATION
56.079 +- 0.038 77.4 +- 1.56 2 AmmountOfDET
55.99 +- 0.142 78.1 +- 3.73 49 POSNPP
```

Annex B: Text collection

author-name	title	year	subgenre	narration
Balzac	Pierrette	1840	Blanche	heterodiegetic
Balzac	TenebreuseAffaire	1841	Policie	heterodiegetic
Balzac	AlbertSavarus	1842	Blanche	heterodiegetic
Sue	MysteresParis02	1842	Fantastique	heterodiegetic
Sue	MorneDiable	1842	Fantastique	heterodiegetic
Sue	MysteresParis01	1842	Fantastique	heterodiegetic

FevalPP	LoupBlanc	1843	Blanche	heterodiegetic
Dumas	Eppstein	1843	Fantastique	heterodiegetic
FevalPP	MysteresLondres1	1843	Policie	heterodiegetic
FevalPP	FanfaronsRoi	1843	Blanche	heterodiegetic
FevalPP	MysteresLondres3	1843	Policie	heterodiegetic
Sue	MysteresParis04	1843	Fantastique	heterodiegetic
Sue	MysteresParis05	1843	Fantastique	heterodiegetic
Sue	JuifErrant	1844	Fantastique	heterodiegetic
Sand	PecheAntoine	1845	Blanche	heterodiegetic
Sue	PaulaMonti	1845	Fantastique	heterodiegetic
FevalPP	Quittance2Galerie	1846	Blanche	heterodiegetic
Sand	LucreziaFloriani	1846	Blanche	homodiegetic
Balzac	CousineBette	1846	Blanche	heterodiegetic
Gautier	PartieCarrée	1848	Fantastique	heterodiegetic
Sue	MysteresPeuple02	1849	Fantastique	heterodiegetic
Dumas	Fantômes	1849	Fantastique	homodiegetic
Dumas	Olifus	1849	Fantastique	homodiegetic
Dumas	ColliersVelours	1850	Fantastique	heterodiegetic
Sue	MysteresPeuple03	1850	Fantastique	heterodiegetic
Sue	MysteresPeuple041850	Fantastique	heterodiegetic	
Sue	MysteresPeuple07	1851	Fantastique	heterodiegetic
Sue	MysteresPeuple06	1851	Fantastique	heterodiegetic
Aurevilly	Ensorcelée	1852	Fantastique	homodiegetic
Ponson	Baronne	1852	Fantastique	heterodiegetic
FevalPP	ReineEpees	1852	Blanche	heterodiegetic
Ponson	FemmeImmortelle	1852	Fantastique	heterodiegetic
Sue	MysteresPeuple09	1853	Fantastique	heterodiegetic
Sue	MysteresPeuple08	1853	Fantastique	heterodiegetic
Sue	MysteresPeuple11	1854	Fantastique	heterodiegetic
Sue	MysteresPeuple10	1854	Fantastique	heterodiegetic
Sue	MysteresPeuple12	1855	Fantastique	heterodiegetic
FevalPP	MadameGilBlas	1856	Blanche	homodiegetic
Gautier	Avatar	1856	Fantastique	heterodiegetic
FevalPP	Louve2	1856	Blanche	heterodiegetic
Sue	MysteresPeuple13	1856	Fantastique	heterodiegetic
Gautier	RomanMomie	1857	Fantastique	heterodiegetic
Sue	MysteresPeuple16	1857	Fantastique	heterodiegetic
Dumas	MeneurLoups	1857	Fantastique	heterodiegetic
Sue	MysteresPeuple15	1857	Fantastique	heterodiegetic
Ponson	ClubValets2	1858	Policie	heterodiegetic
Ponson	ExploitsRocambole3	1859	Policie	heterodiegetic
Ponson	ExploitsRocambole2	1859	Policie	heterodiegetic
Ponson	ExploitsRocambole1	1859	Policie	heterodiegetic
Sand	ElleLui	1859	Blanche	heterodiegetic
Ponson	Chevaliers	1860	Policie	heterodiegetic
Féval	Ténèbre	1860	Fantastique	heterodiegetic
FevalPP	ChevalierTenebre	1861	Fantastique	homodiegetic
Aimard	RodeursFrontieres	1861	Blanche	heterodiegetic
Hugo	Miserables1Fantine	1862	Blanche	heterodiegetic
Ponson	TestamentGrainDeSel	1862	Policie	heterodiegetic
About	OreilleCassée	1862	Fantastique	heterodiegetic
Villiers	Isis	1862	Fantastique	heterodiegetic

FevalPP	HabitsNoirs1	1863	Policie	heterodiegetic
Aurevilly	PrêtreMarié	1864	Fantastique	homodiegetic
Féval	Vampire	1865	Fantastique	homodiegetic
Gaboriau	Lerouge	1865	Policie	heterodiegetic
FevalPP	HabitsNoirs2Coeur	1865	Policie	heterodiegetic
Ponson	Breda	1866	Fantastique	heterodiegetic
Ponson	ResurrectionRocambole2	1866	Policie	heterodiegetic
Ver Ne	CapitaineHatteras	1866	Blanche	heterodiegetic
Ponson	DernierMot3	1867	Policie	heterodiegetic
Ponson	DernierMot4	1867	Policie	heterodiegetic
Gaboriau	EsclavesParis2	1867	Policie	heterodiegetic
Ponson	DernierMot2	1867	Policie	heterodiegetic
Ponson	MiseresLondres3	1868	Policie	heterodiegetic
Aimard	Ourson	1868	Blanche	heterodiegetic
Ponson	MiseresLondres2	1868	Policie	heterodiegetic
Ponson	MiseresLondres4	1868	Policie	heterodiegetic
FevalPP	HabitsNoirs3Rue	1868	Policie	heterodiegetic
Ponson	FéeAuteuil	1868	Fantastique	heterodiegetic
Flaubert	Education	1869	Blanche	heterodiegetic
FevalPP	HabitsNoirs4Arme	1869	Policie	heterodiegetic
FevalPP	HabitsNoirs5Maman	1869	Policie	heterodiegetic
Gouraud	EnfantsFerme	1869	Blanche	heterodiegetic
Gaboriau	MonsieurLecoq2	1869	Policie	heterodiegetic
Zola	FortuneRougon	1870	Blanche	heterodiegetic
Ponson	CordePendu1	1870	Policie	heterodiegetic
Ponson	CordePendu2	1870	Policie	heterodiegetic
Gaboriau	VieInfernale2	1870	Policie	heterodiegetic
Gaboriau	Degringolade1	1872	Policie	heterodiegetic
Gaboriau	Degringolade3	1872	Policie	heterodiegetic
Gaboriau	Degringolade2	1872	Policie	heterodiegetic
Gaboriau	CordeCou2	1873	Policie	heterodiegetic
Zola	VentreParis	1873	Blanche	heterodiegetic
Gaboriau	CordeCou1	1873	Policie	heterodiegetic
Gaboriau	Argent1	1874	Policie	heterodiegetic
Gaboriau	Argent2	1874	Policie	heterodiegetic
FevalPP	VilleVampire	1875	Fantastique	homodiegetic
Zola	AbbeMouret	1875	Blanche	heterodiegetic
Ver Ne	HectorServadac	1877	Fantastique	heterodiegetic
Malot	Cara	1878	Blanche	heterodiegetic
Aimard-Auriac	AigleNoirDacotahs	1878	Blanche	heterodiegetic
Stolz	SecretLaurent	1878	Blanche	heterodiegetic
FevalPP	HommeSansBras	1881	Policie	heterodiegetic
Loti	RomanSpahi	1881	Blanche	heterodiegetic
Boisgobey	Omnibus	1881	Policie	heterodiegetic
Gaboriau	AmoursEmpoisonneuse	1881	Policie	heterodiegetic
Stolz	Mesaventures	1881	Blanche	heterodiegetic
FevalPP	HistoireRevenants	1881	Fantastique	heterodiegetic
Gouraud	ChezGrandMere	1882	Blanche	heterodiegetic
Aurevilly	HistoireSans	1882	Fantastique	heterodiegetic
Maupassant	UneVie	1883	Blanche	heterodiegetic
Rachilde	MVénus	1884	Fantastique	heterodiegetic

Boisgobey	Voilette	1885	Policie	heterodiegetic
Zola	Germinal	1885	Blanche	heterodiegetic
Ohnet	Grande-Marnière	1885	Blanche	heterodiegetic
Zola	Oeuvre	1886	Blanche	heterodiegetic
Villiers	EveFuture	1886	Fantastique	heterodiegetic
Boisgobey	RubisOngle	1886	Policie	heterodiegetic
Malot	Zyte	1886	Blanche	heterodiegetic
Loti	PecheurIslande	1886	Blanche	heterodiegetic
Mary	RogerLa-Honte	1886	Blanche	heterodiegetic
Malot	Conscience	1888	Blanche	heterodiegetic
Boisgobey	OeilChat1	1888	Policie	heterodiegetic
Boisgobey	Chat2	1888	Policie	heterodiegetic
Gouraud	Quand-Grande	1888	Blanche	heterodiegetic
Boisgobey	MainFroide	1889	Blanche	heterodiegetic
Boisgobey	Opera2	1889	Policie	heterodiegetic
Boisgobey	MainFroide	1889	Policie	heterodiegetic
Boisgobey	Opera1	1889	Policie	heterodiegetic
Boisgobey	Double-Blanc	1889	Policie	heterodiegetic

Bibliography

Allison, S., Heuser, R., Jockers, M. L., Moretti, F. and Witmore, M. (2011). *Quantitative Formalism: An Experiment (Stanford Literary Lab, Pamphlet 1)*. Stanford: Standford Literary Lab.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1): 5–32.

Brunner, A. (2015). *Automatische Erkennung von Redewiedergabe: Ein Beitrag Zur Quantitativen Narratologie*. (Narratologia: Contibutions to Narrative Theory Band 47). Berlin; Boston: De Gruyter.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**(3): 1-27.

Cohen, W. W. (1995). Fast Effective Rule Induction. *Twelfth International Conference on Machine Learning*. Morgan Kaufmann, pp. 115–23.

Durrer, S. (1994). *Le dialogue romanesque. Style et structure*. Geneva: Droz.

Elson, D. K. and McKeown, K. R. (2010). Automatic Attribution of Quoted Speech in Literary Narrative. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (AAAI-10).

Genette, G. (1980). *Narrative Discourse. An Essay in Method*. Oxford: Blackwell.

Iosif, E. and Mishra, T. (2014). From Speaker Identification to Affective Analysis: A Multi-Step System from Analyzing Children Stories. *Proceedings of the Third Workshop on Computational Linguistics for Literature*, pp. 40–49.

John, G. H. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, pp. 338–45.

Krestel, R., Bergler, S. and Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. *ELRA, Proceedings of the Sixth International Language Resources and Evaluation Conference*.

Leech, G. N. and Short, M. (1981). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. London; New York: Longman.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit.

Sarmento, L. and Nunes, S. (2009). Automatic Extraction of Quotes and Topics from News Feeds. *Proceedings of DSIE'09 - 4th Doctoral Symposium of Informatics Engineering.*

Weiser, S. and Watrin, P. (2012). Extraction of Unmarked Quotations in Newspapers. *Proceedings of the Eight International Conference on Language Resources and Evaluation.*

#dariahTeach: online teaching, MOOCs and beyond

Susan Schreibman
susan.schreibman@gmail.com
Maynooth University, Ireland

Agiatis Benardou
a.benardou@dcu.gr
Athena Research and Innovation Center in Information Communication & Knowledge Technologies, Greece

Claire Clivaz
claire.clivaz@isb-sib.ch
Swiss Institute of Bioinformatics, Switzerland; University of Lausanne, Switzerland

Matej Ďurčo
matej.durco@oeaw.ac.at
Oesterreichische Akademie der Wissenschaften, Austria

Marianne Huang
mph@au.dk
Aarhus Universitet, Danemark

Eliza Papaki
e.papaki@dcu.gr
Athena Research and Innovation Center in Information Communication & Knowledge Technologies, Greece

Stef Scagliola
scagliola@eshcc.eur.nl
Erasmus Universiteit Rotterdam, Netherlands

Toma Tasovac
ttasovac@humanistika.org
Belgrade Center for Digital Humanities, Serbia

Tanja Wissik
tanja.wissik@oeaw.ac.at
Oesterreichische Akademie der Wissenschaften, Austria

Background

Online education has been advocated as the ultimate way of democratizing knowledge, but recent research indicates that there are reasons for concern. As the Allen & Seaman 2014 report underlines, 66% of higher education institutions report that online education remains critical to their long-term strategy while 74% of chief academic officers consider the learning outcomes for online courses to be 'as good as or better' than traditional face-to-face courses. But "despite this confidence in online education, researchers continue to report 'compromised quality in online courses' as one of the concerns of faculty, administration, and the general public" (Kidder, 2015; Selingo, 2014). In the landscape of online teaching, MOOCs (Massive Open Online Courses) have received much attention in both academic and popular publications (Bayne and Ross, 2015; Bulfin et al., 2014; Clara and Barbera, 2013) despite the fact that they are not representative of the diverse modalities of online teaching.

Siemens (2012) makes a useful distinction between xMOOCs (behaviorist MOOCs) and cMOOCs (connectivist MOOCs). The former emphasizes "a more traditional learning approach through video presentations and short quizzes and testing" with a focus on "knowledge duplication", whereas the latter focus on "knowledge creation" (Siemens, 2012). Along the same lines, Ozturk recently reported that new variations of MOOCs have emerged becoming more market oriented "aligning with instructivist, cognitive, and behaviourist pedagogy" (Ozturk, 2015). Moreover, the financial model of the MOOCs raises questions about the audience for and motivations behind this method of teaching (Ozturk, 2015; Manjoo, 2015).

Conscious of this present situation, the #dariahTeach project (funded by an Erasmus+ Strategic Partnership) is developing a network based in seven partner countries exploring the production, dissemination, and promotion of high quality, dynamic, extensible, localisable, and integrated educational materials for the digital humanities specifically tailored for third level education. It is adopting a cMOOC philosophy which focuses on 'creation, creativity, autonomy, and social networked learning' (Siemens, 2012) to provide pedagogical content that can be easily integrated into diverse teaching and learning situations.

A key consideration in the design of the platform is interoperability between courses/modules (and units within those modules) since DH draws on a wealth of methods and tools from a variety of disciplines. Moreover, it is envisioned that these modules will be used beyond the DH community as the societal impact of a culturally-driven digital transition grows opening up new ways of collaborating on productive theory and critical thinking (Hayles, 2012). Thus a goal of #dariahTeach is to develop rich educational materials that 1) instructors in the growing number of Digital Humanities programmes can use as appropriate to their own institutional settings and learning outcomes; 2) instructors in other disciplines can draw on and; 3) students who are not at institutions that have DH expertise can use to develop the skills and methods, as well as understand the theoretical basis, to engage in digital humanities and humanities research.

The project team is currently developing the infrastructure and design of the modules based on the production of five modules: Introduction to Digital Humanities, Text Encoding, AudioVisual Media and Multimodal Literacies, Retrodigitizing Dictionaries, and Ontologies and Knowledge Management. This paper will present the results of preliminary research carried out through an extensive study of user requirements, as well as desk research on module and platform design informed by a workshop in Belgrade funded by the Digital Research Infrastructure for the Arts and Humanities (DARIAH) on developing open educational materials.

Analysis of User Requirements

The design and the implementation of a successful platform-based learning environment melds concepts from psychology, education, and human-computer interaction. Poor interface design can become a serious obstacle to the learning outcome, as it may slow the process down and impose cognitive obstacles. To this end, a qualitative analysis and interpretation of online teaching practices and recommendations in the DH domain and the elicitation of corresponding user requirements was based on a series of semi-structured interviews with experienced instructors of online courses within Europe.

Findings of the user requirement process are a key component of the development of the #dariahTeach platform. These indicate that the platform needs to cater for the following needs: be adaptable to different learning methodologies; allow for persistent roles; provide an API or advanced forms of web services so that new unforeseen components can be added to the environment; support ad hoc groupings and grouping of materials across modules and units; allow for both synchronous and asynchronous collaboration and communication and enable user customization.

Module Design

#dariahTeach modules are designed as building blocks tailored to the exigencies of teaching situations in different educational and cultural contexts, allowing for localization and adaptation (via translation, subtitles, domain-specific examples etc.). By offering examples of and encouraging further adaptation of training materials to specific linguistic/cultural contexts, #dariahTeach will dispel any notion that the use of ICT methods leads to abstract representations of culturally impoverished outputs.

It is important to stress two levels of translatability of module design: 1) translatability and adaptability of the language of instruction; and b) selectability, translatability and adaptability of primary sources and materials that are used in instruction. This means that an English-language module on Text Encoding, for instance, is localizable both in terms of the instructional narrative, as well as the kind of texts that are used to exemplify the taught principles and methods of text modeling: different genres (poetry, prose, drama) but also language (Latin, Greek, Serbian, Dutch etc.)

Our "Introduction to DH" module will also not attempt to impose a single pedagogical narrative on what is a constantly evolving and highly diverse, interdisciplinary field. Instead, our Introduction to DH is based on a micromodular, polycentric approach: a collection of mutually-linked, cross-referenced, metadata-rich short videos that shed light on DH as a community of practice from multiple perspectives without creating a false sense of uniformity.

Platform Design

Modules will be made available via an online portal/web application based on existing solutions. This paper will explore the decision tree in adopting a solution including whether to use a well-established Content Management Systems (eg Drupal, WordPress, Joomla) with Learning Management System plugins and appropriate customizations or the use of a customizable Learning Management System (such as Moodle or Blackboard). Considerations feeding into the decision tree include the platform being open source, freely available, well documented and customizable with plugin development support; support for multilinguality; an embedded xml editor; collaboration and interaction functionalities (eg chats, forums and wikis); test and assessment functionalities; extended search functionalities for available metadata (mapped to Dublin Core and LOM to facilitate sharing and support interoperability and reusability (Roy et al., 2010); and copyright attribution and licence management functionalities.

Conclusion

The paper will conclude with longer-term prospects for the project. Oversight of #dariahTeach will be maintained after the grant has ended by a General Editor and Editorial Board under the oversight of the DARAH's Research and Education Competency Centre.

Bibliography

Allen, I. E. and Seaman, J. (2014). *Grade Change – Tracking Online Education in the United States,* Babson Survey Research Group and the Sloan Consortium, LLC. http://www.onlinelearningsurvey.com/reports/gradechange.pdf

Bayne, S. and Ross, J. (2015). MOOC Pedagogy. In Kim, P. (ed.) *Massive Open Online Courses: The MOOC Revolution*. Oxford: Routledge.

Bulfin, S., Pangrazio, L. and Selwyn, N. (2014). Making 'MOOCs': The construction of a new higher education within news media discourse. *International Review of Research in Open and Distance Learning*, **15**(5): 209-305.

Clarà, M. and Barberà, E. (2013). Learning online: massive open online courses (MOOCs), connectivism, and cultural psychology. *Distance Education* **34**(1): 129-36.

Ferguson, R. and Sharples, M. (2014). *Innovative Pedagogy at Massive Scale: Teaching and Learning in MOOCs. Open Learning and Teaching in Educational Communities*. Springer.

Hayles, N. K. (2012). *How We Think. Digital Media and Contemporary Technogenesis*. Chicago University Press.

Kidder, L. C. (2015). The Multifaceted Endeavor of Online Teaching: The Need for a New Lens". In Hokanson, B., Clinton, G., Tracey, M. (Eds.) *The Design of Learning Experience Creating the Future of Educational Technology*. Springer, pp. 77-91.

Manjoo, F. (2015, 16 September). 'Udacity Says It Can Teach Tech Skills to Millions, and Fast'. *The New York Times*, http://nyti.ms/1ihbcp7

Ozturk, H. T. (2015).Examining Value Change in MOOCs in the Scope of Connectivism and Open Educational Resources Movement". *International Review of Research in Open and Distributed Learning 16/5, Creative Commons 4.0.*

Peters, D. (2014). *Interface Design for Learning: Design Strategies for Learning Experiences*. San Francisco: New Riders.

Roy, D., Sarkar S. and Ghose S. (2010). A Comparative Study of Learning Object Metadata, Learning Material Repositories, Metadata Annotation and an Automatic Metadata Annotation Tool. In Joshi, M., Boley, H., Akerkar, R. (eds.). *Advances in Semantic Computing* **2**: 103-26.

Selingo, J. J. (2014). "Demystifying the MOOC". *The New York Times*, http://nyti.ms/1u6MYCL

Siemens, G. (2012). MOOCs are really a platform, In idem *Elearnspace blog*, http://www.elearnspace.org/blog/2012/07/25/moocs-are-really-a-platform/

The Computer Graphic Simulation of the Battle at Mount Street Bridge. Problems, Perspectives, and Challenges

Susan Schreibman
susan.schreibman@gmail.com
An Foras Feasa, Maynooth University, Ireland

John Buckley
John.Buckley@iadt.ie
Department of Film and Media, Dun Laoghaire Institute of Art, Design and Technology

Brian Hughes
B.Hughes@exeter.ac.uk
Department of History, University of Exeter

Constantinos Papadopoulos
cpapadopoulos84@gmail.com
An Foras Feasa, Maynooth University, Ireland

Introduction

Three-dimensional digital visualisations have been used in the last three decades to engage and educate by making complex data more comprehensible. These have been used across a wide range of fields including military, medicine, cultural heritage, archaeology, and history. In cultural heritage contexts, the potential of these technologies as tools in the process of knowledge production has been well demonstrated (see for example Sundstedt et al., 2004; Papadopoulos et al., 2015). Contested Memories: The Battle of Mount Street Bridge project has been exploring how to integrate a 3D visualisation as the primary text of a digital scholarly edition, raising issues of how the phenomenology of place and space can be used to design a new language of scholarly editions, one that has the ability to model experience lost because of technological and evidentiary constraints. This edition, like traditional DSEs, also brings together documentary evidence in the form of apparatus, reimagining digital textuality (see Snyder, 2015).

The project focuses on a battle that took place on Wednesday 26th April 1916 during the week of the Easter uprising in Dublin. This particular battle, between a small group of Irish rebels (members of the Irish Volunteers) and the British army sent to Dublin to put down the rebellion is used here to investigate to what extent a networked virtual world can be accommodated into a scholarly editing paradigm developed for print (and more recently digital) to enable alternative forms of research, help in the interpretive process, and assist knowledge production for both general audiences and specialists.

The development of the primary text in a gaming platform posed significant challenges, many of which have been ameliorated when creating digital editions of print or manuscript sources through such methods and standards as XML, The Text Encoding Initiative, and XML-aware databases. This case study will explore these challenges, the project's accomplishments, and future directions.

Contested Memories: The Battle of Mount Street Bridge is one of the four projects developed under the Humanities Virtual World Consortium (HVWC) and is funded by The Andrew W. Mellon Foundation. Four virtual worlds are being constructed, each released as Unity3D builds (2015), while a shared virtual world platform ensures the sustainability and interoperability of these as well as future projects.

Historical Background and Purpose of the Digital Simulation

The Battle of Mount Street Bridge has attracted much scholarly attention not only due to its significance during the Easter Rising but also because of the varied and contradictory accounts of British casualties, as well as the timeline and British strategy during the battle. Seventeen

Irish volunteers took positions at four different buildings in and around on a one block stretch along Northumberland Road in a leafy suburb of Dublin (fig. 1). They prevented two Battalions of British soldiers known as the Sherwood Foresters (some 1,750 men) for over six hours from progressing into the city centre.

Although there exists a significant amount of documentation for the battle from both British and Irish sources, some of it written in the 1920s (The Robin Hoods, 1921; Oates, 1920), with other sources given as oral histories between 1945-1955 (see the Bureau of Military History), scholars have not been able to map, both spatially and temporally, the contours of the engagement so as to understand how such a small number of rebels could inflict such heavy losses on a trained group of soldiers.

Thus, a goal of the project is to employ spatiotemporal methods in order to create a model of the environment in which the battle took place as well as simulate certain key events to enable researchers and specialists to visualise the conflict as well as to understand how the conflict unfolded.

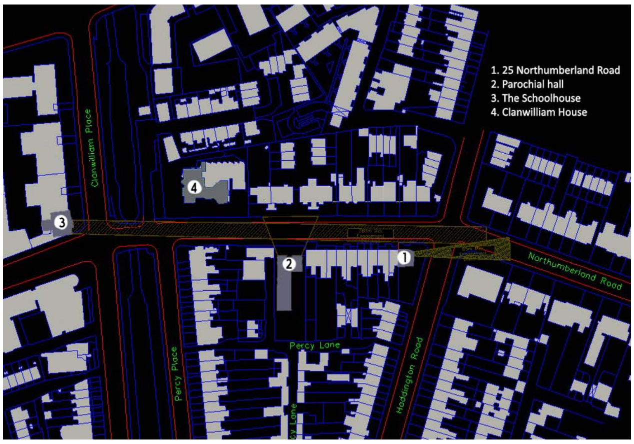


Figure 1. Location of buildings occupied by the Irish Volunteers

Methodology

This project employed a range of research methods in order to obtain relevant information for the battle and inform the decision-making process of building the virtual world. Firstly, conventional archival research and meetings with military historians were carried out to document different sources (both contradictory and fragmentary) that provide evidence for the buildings that were occupied, participant accounts, the weapons they used, as well as the events that took place during the battle. Period photographs were used for the digital (re)construction of the area, while a Lidar Scanning of Northumberland Road helped in creating a highly accurate digital model of the battle scene. Ballistic experts and experiments at the shooting range provided a well-informed recreation of bullets' trajectories and guns' sounds and reload rates.

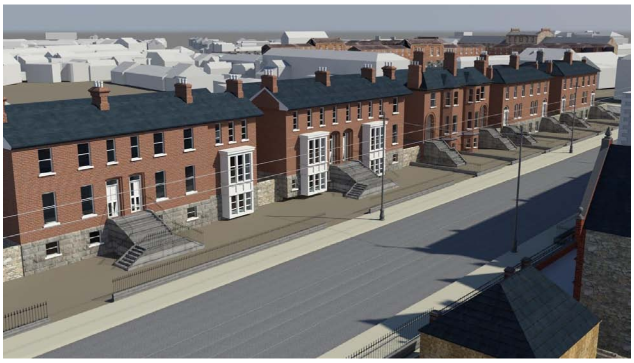


Figure 2. A 3D Model of Northumberland Road in 3ds Max 2015.

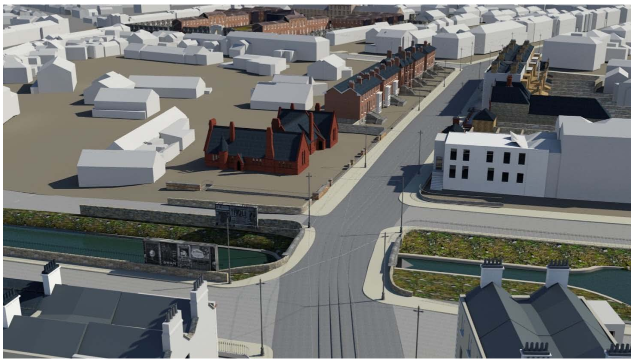
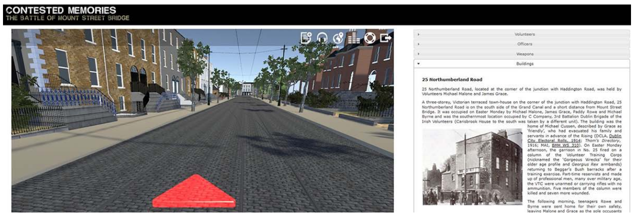


Figure 3. Rendering of the battlefield. Building and features that were significant in the conflict were modelled in more detail, while a schematic view of the broader area provides adequate context to the users.

The above information was used as the basis to construct a 3D model of the battle scene, as well as annotations for the Irish Volunteers and the British forces, weapons, and buildings at Northumberland Road and the adjacent streets. The purpose of this visualisation is not to present an exact representation of the battle. Rather, it is a tool to investigate alternative interpretations, as well as provide a case study to engage in ongoing interdisciplinary debates regarding the nature of digital reconstructions (Bentkowska-Kafel et al. 2012; Clark 2010; Kensek et al. 2004). The 3D model was constructed in 3ds Max 2015 (Autodesk 2015) (fig. 2, 3) and was migrated to the game engine Unity3D (2015) in order to enable a navigable in-browser 3D world for users to explore. The first version of the virtual world (November 2015) used all the capabilities of Unity Web Player (fig. 4). However, given the limited support of NPAPI, the plugin framework that Unity Web Player uses to enable detailed and high resolution models to properly function in web browsers, the second version (February 2016) implemented WebGL technology, which at the time of writing this paper, only supports light-weight applications. Therefore, most models of the virtual world had to be optimised, while detailed models of the city that provide contextual information for the Battle had to be omitted (fig. 5). Users of the virtual environment are able to navigate the world in first person-mode, while not having in-world representation in order to avoid modern figures being visible in the battle field.



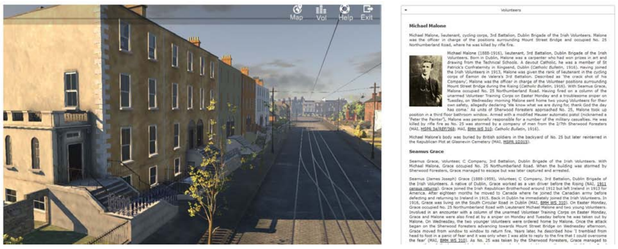
25 Northumberland Road

25 Northumberland Road, located at the corner of the junction with Haddington Road, was held by Volunteers Michael Malone and James Grace.

A three-storey, Victorian terraced town-house on the corner of the junction with Haddington Road, 25 Northumberland Road is on the south side of the Grand Canal and a short distance from Mount Street Bridge. It was occupied on Easter Monday by Michael Malone, James Grace, Paddy Rowe and Michael Byrne and was the southernmost location occupied by C Company, 3rd Battalion Dublin Brigade of the Irish Volunteers (Carisbrook House to the south was taken by a different unit). The building was the home of Michael Cussen, described by Grace as 'friendly', who had evacuated his family and servants in advance of the Rising (DCLA, Dublin City Electoral Rolls, 1914; *Thom's Directory*, 1916; MAI, BMH WS 310). On Easter Monday afternoon, the garrison in No. 25 fired on a column of the Volunteer Training Corps (nicknamed the 'Gorgeous Wrecks' for their older age profile and *Georgius Rex* armbands) returning to Beggar's Bush barracks after a training exercise. Part-time reservists and made up of professional men, many over military age, the VTC were unarmed or carrying rifles with no ammunition. Five members of the column were killed and seven more wounded.

The following morning, teenagers Rowe and Byrne were sent home for their own safety, leaving Malone and Grace as the sole occupants

Figure 4. First version of the Virtual World embedded within the digital scholarly edition



Michael Malone

Michael Malone, lieutenant, cycling corps, 3rd Battalion, Dublin Brigade of the Irish Volunteers. Malone was the officer in charge of the positions surrounding Mount Street Bridge and occupied No. 25 Northumberland Road, where he was killed by rifle fire.

Michael Malone (1888-1916), lieutenant, 3rd Battalion, Dublin Brigade of the Irish Volunteers. Born in Dublin, Malone was a carpenter who had won prizes in art and drawing from the Technical Schools. A devout Catholic, he was a member of St Patrick's Confraternity in Ringsend, Dublin (*Catholic Bulletin*, 1916). Having joined the Irish Volunteers in 1913, Malone was given the rank of lieutenant in the cycling corps of Éamon de Valera's 3rd Battalion. Described as 'the crack shot of his Company', Malone was the officer in charge of the Volunteer positions surrounding Mount Street Bridge during the Rising (*Catholic Bulletin*, 1916). With Seamus Grace, Malone occupied No. 25 Northumberland Road. Having fired on a column of the unarmed Volunteer Training Corps on Easter Monday and a troublesome sniper on Tuesday, on Wednesday morning Malone sent home two young Volunteers for their own safety, allegedly declaring 'We know what we are dying for, thank God the day has come.' As units of Sherwood Foresters approached No. 25, Malone took up position in a third floor bathroom window. Armed with a modified Mauser automatic pistol (nicknamed a 'Peter the Painter'), Malone was personally responsible for a number of the military casualties. He was killed by rifle fire as No. 25 was stormed by a company of men from the 2/7th Sherwood Foresters (MAI, MSPR 34/REF/368; MAI, BMH WS 310; *Catholic Bulletin*, 1916).

Michael Malone's body was buried by British soldiers in the backyard of No. 25 but later reinterred in the Republican Plot at Glasnevin Cemetery (MAI, MSPR 1D315).

Seamus Grace

Seamus Grace, Volunteer, C Company, 3rd Battalion, Dublin Brigade of the Irish Volunteers. With Michael Malone, Grace occupied No. 25 Northumberland Road. When the building was stormed by Sherwood Foresters, Grace managed to escape but was later captured and arrested.

Seamus (James Joseph) Grace (1888-1959), Volunteer, C Company, 3rd Battalion, Dublin Brigade of the Irish Volunteers. A native of Dublin, Grace worked as a van driver before the Rising (NAI, 1911 census returns). Grace joined the Irish Republican Brotherhood around 1912 but left Ireland in 1913 for America. After eighteen months he moved to Canada where he joined the Canadian army before defecting and returning to Ireland in 1915. Back in Dublin he immediately joined the Irish Volunteers. In 1916, Grace was living on the South Circular Road in Dublin (MAI, BMH WS 310). On Easter Monday, Grace occupied No. 25 Northumberland Road with Lieutenant Michael Malone and two young Volunteers. Involved in an encounter with a column of the unarmed Volunteer Training Corps on Easter Monday, Grace and Malone were also fired at by a sniper on Monday and Tuesday before he was taken out by Malone. On Wednesday, the two younger Volunteers were ordered home by Malone. Once the attack began on the Sherwood Foresters advancing towards Mount Street Bridge on Wednesday afternoon, Grace moved from window to window to return fire. Years later, he described how 'I trembled from head to foot in a panic of fear and it was only when I was able to reply to the fire that I could overcome the fear' (MAI, BMH WS 310). As No. 25 was taken by the Sherwood Foresters, Grace managed to

Figure 5. Second version of the Virtual World in WebGL Unity Web Player

Conclusion

The digital simulation of the Battle of Mount Street Bridge provides a novel methodology for knowledge production and understanding in historical research demonstrating how computer-based simulations can augment traditional approaches in historical datasets, enhance the interpretive process, and potentially provide answers in complicated research questions. It was the process of producing the digital simulation and not simply the end-product that provided valuable answers to our questions. Once the model was completed, however, new research questions emerged, as historians interested in the period, as well as the public, began interacting with the model.

Future development of the project includes embedding narratives about in-world objects (buildings, avatars etc.), contextual information and decisions, animations of the broad contours of the battle, an AI simulation of key events, and a Mixed-Reality application tailored to the general public and secondary school students.

Bibliography

3ds Max, (2015). *Autodesk*. http://www.autodesk.com/products/3ds-max/ (accessed 27 February 2016).

Bentkowska-Kafel, A., Denard, H. and Baker, D., (Eds.) (2012). P *aradata and Transparency in Virtual Heritage*. England, UK: Ashgate

Bureau of Military History: Accounts/Documents/Images/Audio (1913-1921). Defence Forces Ireland. http://www.bureauofmilitaryhistory.ie/ (accessed 27 February 2016).

Clark, J. T. (2010). The Fallacy of Reconstruction. In Forte, M. (Ed) *Cyber-archaeology*. Oxford: Archaeopress.

Kensek, K., Dodd, L., Cipolla, N. (2004). Fantastic Reconstructions or Reconstructions of the Fantastic? Tracking and Presenting Ambiguity, Alternatives, and Documentation in Virtual Worlds. *Automation in Construction*, **13**: 175–86.

Oates, W. C. (1920). The Sherwood Foresters in the Great War: 1914-1918. *The 2/8th Battalion*. Nottingham: J and H Bell.

Papadopoulos, C., Hamilakis, Y. and Kyparissi-Apostolika, N. (2015). Light in a Neolithic dwelling: Building 1 at Koutroulou Magoula (Greece). *Antiquity*, **89**(347): 1034-50.

Snyder, L. (2015). VSim Software. Institute for Digital Research and Education, UCLA. https://idre.ucla.edu/research/active-research/vsim (accessed 27 February 2016).

Sunstedt, V., Chalmers, A. and Martinez, P. (2004). High Fidelity Reconstruction of the Ancient Egyptian Temple of Kalabsha. *AFRIGRAPH '04 Proceedings of the 3rd International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, pp. 107–13. New York: ACM.

The Robin Hoods: 1/7th, 2/7th and 3/7th Battns. Sherwood Foresters 1914-1918 (1921). Written by the Officers of the Battalions. Nottingham: J&H Bell, Ltd.

Unity3D Game Engine (2015). Unity. Version 5. http://unity3d.com/ (accessed 27 February 2016).

The Mutual Relationship of Linguistic and Non-linguistic Elements in Breaking Down the Hierarchy of Language in Digital Narrative

Mehdy Sedaghat Payam
ms79payam@yahoo.com
SAMT Center for Research and Development in Humanities

Although there have been a handful of transmedial explorations by some novelists in the past, narrative in novels has mainly used linguistic elements in conventional print fiction. When the first generation of hypertext novelists remediated print novels in their works they only used language too. However, images, sound, movies and other non-linguistic elements in works of the second generation digital novels—called web-fiction here—have mainly challenged the hierarchy of language and have posed serious threats to the autonomy of words in novels. This challenge has worked both ways, on one hand, works of web-fiction have tried to distance themselves from works which have been merely linguistic (for instance *Afternoon* (1990)), on the other hand experimental print novelists have become media-conscious and created novels which have incorporated images, colors, innovative page design and music in their physical body (for instance *House of Leaves* (200)). Consequently these media-conscious novels, especially in digital media, have posed serious questions for novel. Some of these questions are: How much narrative in novels is dependent upon words and in what ways can it take advantage of the narrative potentials of the non-linguistic elements? In other words how can non-linguistic elements contribute to the narrative of novels in the media in which they are rendered and in what ways does this new synthesis of words and non-linguistic elements can change our understanding of the narrative in novels? Finally how these novels should be analyzed? In order to find the answers of these questions, two works of web-fiction, *Reconstructing Mayakovsky* (2011) by Illya Szilak and *Dreamaphage* (2006) by Jason Nelson, and one work of media-conscious print *S* (2013) by Doug Dorst and J.J. Abrams have been chosen and the mutual relationship between their linguistic and non-linguistic elements have been explored. All of these works have tried to break down the hierarchy of language in the narrative of the novel and in doing that have highlighted the role of non-linguistic features and have highlighted the ways these features can contribute to narrative in a novel especially in a digital medium.

Analyzing these works will help us to find the answer to a bigger question. How do these mixtures of several media justify their existences as novels? This is where pushing Bakhtin's ideas a little bit further to include works of digital fiction can become extremely useful. Bakhtin's ideas have been used as the theoretical base of my discussion because out of the three other theories of novel presented in the *Routledge Encyclopaedia of Narrative* (2005) , it is the only one which takes the materialities of the production of the text into consideration and according to Howard Mancing "seems most justified by an informed understanding of literary history and theory" (ibid, 399). According to Bakhtin "There is no specific form, technique, theme, or approach to character that makes a text a novel; rather, the distinguishing characteristics of the novel are its heteroglossia and its dialogism" (ibid, 400). Since Bakhtin does not limit dialogism to literature only and believes all language (and all thought) is dialogical, in this paper it has been tried to extend these two concepts into "modes" in both digital and media-conscious print novels. In these novels multiplicity of modes can result in novels in which the words' hold over the narrative is not as strong as it has been in conventional novels. Such novels can potentially provide a dialogical engagement between linguistic and non-linguistic elements which can eventually lead to a different understanding of what novel is (or it can be). In this way, the current proposal aims at providing a theoretical background and justification for these kinds of narratives which claim to be novels and offers a practical method to read and analyze these novels.

Historical Background

The introduction of digital media to the literary scene encouraged a number of experimental novelists—some of whom like Joyce had published novels in print—to try their hand at this medium to create works which were both written on the computer and necessarily had to be read on the screen as well. These writers who were later known as hypertext novelists experimented with the materiality of this new medium, and made its materiality an explicit part of the conception of their novels. Although some conventions of the print medium were discernible in the works of these writers, the arrival of the Internet and the developments in digital media provided a significant opportunity for these writers and the new generation of writers to experiment with new conventions for novel writing in digital media. Moreover, these experimentations with the materiality of the medium encouraged the experimental print novelists of the digital era to experiment more extensively with print as a medium.

N. Katherine Hayles is a prominent scholar who has consistently written on the "materiality of the medium" and this paper has heavily borrowed from her theoretical discussions and coinages. Hayles advocates a method of reading called Media-Specific Analysis (MSA) which involves paying particular attention to the materiality of the medium in which the work of fiction is presented. The importance of Hayles' analytical method is that it provides a practical method for thinking about text as a linguistic object, and provides a new perspective to think and write about texts. Another advantage of Hayles' approach is that it brings the medium to the foreground from the very beginning and can be applied for the analysis of both print and digital novels.

Since the novel as a specific genre and media form developed its defining characteristics and conventions in association with the evolution of print technology, the question of how narrative in novels is transformed through works of hypertext or web-fiction is a significant one. H. Porter Abbott has a useful definition of narrative and his definition will be used as a guide in the controversial subject of narrative and how it should be thought of in the works of hypertext and web-fiction. In Abbott's definition, "narrative is the representation of events, consisting of *story* and *narrative discourse.*" Story "is an event or sequence of events (the action), and narrative discourse is those events as represented" (16). The main reason that Abbott's definition has been chosen here is that it can be applied to the study of narrative in an almost any medium.

The non-linguistic elements which have been used in media-conscious novels are referred to as modes here, therefore a novel which has used several modes in its narrative is a multimodal novel. This usage of the term mode is more in line with the way Alison Gibbons has defined this term as "a system of choices used to communicate meaning." Looking at novel from the modality perspective, provides us a better understanding of how each work is created out of the different combinations of modalities of three different but related categories. A fictional text uses a specific modality of the text which is the narrative genre. It uses the modality of the medium either print or digital. The last modality which comes into play here is the modality of verbal/visual which is part of the modalities of representation. These modalities can work in different combinations, but segregating them in this way, makes them more visible and shows how each writer can create texts, by manipulating either of these so that the reading process would be affected by the way either of these modalities is brought into play.

Thus, in order to study the he fictional works discussed here, three different but interrelated dimensions of the fictional text have been identified. (1) Physical Organization and Design, (2) Narrative Strategy, and (3) Reading Process. This tripartite model can be used by other scholars for analyzing novels which incorporate linguistic and non-linguistic elements in their narrative(s). In the first dimension, Physical Organization and Design a text is analyzed from the perspective of the use of its physical resources its authorship, and design. Narrative Strategy, the second dimension, is the angle through which a text from the perspective of the use of its physical resources and signifying strategies to create a narrative is analyzed. In the third dimension, the Reading Process, the way a text shapes the experience of the reader is explored.

Novels have always been media forms which lead the reader through them to a world which is the real world or like the real world in its spatial/visual form. These worlds exist beyond the page and the language and materiality of the novel are expected to be effaced during this process. However, the digital novels and media-conscious print novels show resistance toward this self-effacement and by mixing different modes in their narrative and bringing their own materiality into the foreground. The narrative and how it is created through the interplay of different modes in a single work is analyzed in these three novels:

- *Reconstructing Mayakovsky* by Illya Szilak. This novel which claims to be the "novel of future" has been published in electronic format in the second volume of ELO in 2011. Later in 2012, a print version of the same novel was published by Revolution Nostalgia Disco Theater. Mayakovsky's believed that poetry is a mode which can disrupt a fixed narrative and this is exactly what Szilak tries to accomplish in her novel. The narrative in this novel is broken into several modes—or "mechanisms," as the writer calls them—including text, audio podcasts, video, a live Google image search, etc.
- *Dreamaphage* by Jason Nelson. The interfaces of Dreamaphage have been described as "innovative, surprising, alternately whimsical and unnerving" by ELO website and these words can be used to describe the whole narrative of novel as well. Narrative in this novel is presented through different layers of 3D text with several links to the books which the reader has to simulate the act of turning the page (which means clicking the mouse on the bottom right side of the screen where the edge of the book is and move it to the top left, so that the next page appears on screen), and move through pages filled with squares and circle to read some files which their loose connection together forms a narrative.
- *S* by Doug Dorst and J.J. Abrams. This print novel comes in a slip cover which has a book named "Ship of Theseus" which is full of handwritten comments, as well as postcards, newspaper clippings, black and white photos, even a hand-drawn map written on a napkin from a coffee shop, which together create a mysterious narrative. These non-linguistic materials play such a crucial role in the narrative, that at times it becomes impossible to follow the narrative and solve the mystery of the novel without them. *S* is described a "love letter to the written word," on its slipcover, however it is through the interaction of the linguistic and non-linguistic elements that this love letter finds its true meaning.

Bibliography

Abbott, H. Porter (2002). *The Cambridge Introduction to Narrative*. Cambridge: Cambridge University Press.

Bakhtin, M. (1981). *Dialogic Imagination.* Austin: University of Texas Press.

Danielewski, M. Z. (2000). *House of Leaves*. New York: Pantheon Books.

Dorst, D. and Abrams, J. J. (2013). *S*. New York: Mulholland Books.

Hayles, N. K. (2004). Print Is Flat, Code Is Deep: The Importance of Media-Specific Analysis. *Poetics Today*, **25**(1): 67–90.

Herman, D., Manfred J., and Ryan, M.-L. (eds.) (2005). *Routledge Encyclopedia of Narrative Theory*. London; New York: Routledge.

Joyce, M. (1990). *Afternoon; a story*. Watertown, MA: Eastgate Systems.

Nelson, J. (2006). *Dreamaphage*. Electronic Literature Collection Volume **1**.

Szilak, I. (2011). *Reconstructing Mayakovsky*. Electronic Literature Collection Volume **2**.

What Do Boy Bands Tell Us About Disasters? The Social Media Response to the Nepal Earthquake

David Lawrence Shepard
shepard.david@gmail.com
UCLA, United States of America

Takako Hashimoto
takako@cuc.ac.jp
Chiba University of Commerce, Japan

Tetsuji Kuboyama
kuboyama@tk.cc.gakushuin.ac.jp
Gakushuin University, Japan

Kilho Shin
yshin@ai.u-hyogo.ac.jp
Hyogo University, Japan

When an earthquake struck Nepal in 2015, the band One Direction sent a tweet encouraging their fans to donate to relief efforts. This one tweet was retweeted a few times, but quickly lost in a flood of other tweets about One Direction's tour. Simultaneously, an Indian Hindu extremist politician flooded his Twitter stream with rumors that Christian missionaries were coercing conversions from Nepalis in exchange for humanitarian aid. Additionally, an Indian religious group mixed substantial numbers of tweets about a movie they had released with tweets about their relief efforts in Nepal. These are just a few of the users who engaged with the disaster from a distance: they had different motives for tweeting about the disaster, and different levels of engagement with it. We call these users "onlookers:" they tweet about a disaster, but are not directly affected by it.

This paper analyzes onlooker behavior: it argues that onlookers who will send a few tweets can be predicted by their interests, while onlookers who will tweet heavily about it have few, if any, shared interests. We show that onlookers who primarily tweet about entertainment topics and news topics are likely to mention the disaster, yet send few tweets about it. On the other hand, onlookers who tweet substantially about a disaster after it happens are difficult to identify before the disaster occurs because they do not share common interests aside from the disaster.

Background

Natural disasters often attract significant attention on Twitter, both by those affected and those who are distant. A substantial amount of research has explored how social media causes users to engage with political, social, and humanitarian problems; however, opinions on social media's effectiveness—whether it causes users to donate money, stay informed, or participate in campaigns—are mixed. Some argue that displaying concern in social media is more about acquiring social capital than effecting change (Shulman, 2009; Gladwell, 2011; Morozov, 2012; Morozov, 2014), while a Pew Research Center survey finds that social media does create change (Raine et al., 2016). Additionally, many have argued that social media was important though not essential to protests in Egypt (Mazaid, 2011; Tufekci and Wilson, 2012) and other nations (Raine et al., 2016; Shirky, 2011). One analysis found that charities' use of social media does not increase donations (Malcolm, 2016), while another finds that certain tweeting strategies do (Gasso Climent, 2015) although tweets may not raise awareness about the charity's causes (Bravo and Hoffman-Goetz, 2015). Where all these studies concur is that social media enable a substantial amount of discourse about crises. The question we explore is how to predict how much attention Twitter users pay to crises: social media presents the opportunity for a user to send a single retweet about a disaster—as many One Direction fans did—or to sustain interest by following other users and sending many tweets about the event over a period of time.

Additionally, there is little question that social media is useful for those directly affected by disasters. A substantial amount of research finds that social media helps first responders (Regalado et al., 2015; Dugdale et al., 2012; Omilion-Hodges and McClain, 2016; Burns, 2015; Xiao et al., 2015; Kaewkitipong et al., 2016; Meng et al., 2015; Madianou, 2015; Palen, 2008). In fact, specialized algorithms have been developed for that purpose (Pohl et al., 2013a; Pohl et al., 2013b; Platt et al., 2011). Little work examines users who tweet about disasters at a distance, however, despite the large numbers of such users. We examine these onlookers because they produce a large amount of the tweets about humanitarian crises.

Method

We use quantitative text analysis to identify tweets about the earthquake, to cluster onlookers based on shared interests, and to derive a correlation between onlookers' interests and the number of tweets they sent about the earthquake. This section outlines our methods.

To attain a broad sample of onlookers, we gathered a dataset of over 5 million tweets sent by around 15,000 users in the three weeks following the Nepal earthquake. We harvested the data from the Twitter REST API by searching for any tweets that mentioned the word "Nepal" from April 24 to May 8. We randomly selected 15,000 users from this set and harvested all of their tweets sent between April 24, 2014 and May 5, 2015. We attempted to capture only English-speaking users to increase the likelihood that we would capture users not directly affected by the earthquake, but we still captured some users who tweeted in multiple languages. This left us with roughly 11,000 onlookers.

To determine how many tweets a user had sent about the earthquake, we trained a Naive Bayesian classifier using MALLET (McCallum, 2002) on a set of 100 onlookers' tweets (totaling about 30,000 tweets), marking them as either quake-related or not. We applied the trained classifier to the remainder of the dataset to count each user's quake-related tweets. Spot checking showed this technique had acceptable accuracy.

To find shared interests, we used Latent Dirichlet Allocation (Blei et al., 2003), treating all of a user's tweets as one document. We ran LDA with MALLET with various numbers of topics, and settled on 80. These topics represented a broad span of themes: greetings, news, entertainment, technology, plus four topics directly related to the earthquake. We then looked for connections between onlookers by building an edge list of shared topics, creating a weighted edge between two onlookers if over 25% of both onlookers' tweets consisted of a shared topic. The edge weight was the product of their affinities to that topic. Using Gephi (Bastian et al., 2009), we then ran a weighted Louvian modularity algorithm (Blondel et al., 2008) over this onlooker graph to generate communities of users.

Analysis

This experiment resulted in 21 communities of onlookers being identified. The communities were labelled using the strongest topics in each.

ID	Label	Average Number of Tweets	Users	Average Quake-Related Tweets
0	Foreign language (Spanish)	419	882	9
1	Japanese Music	403	135	5
2	Greetings	326	199	5
3	Portuguese/Fifth Harmony	710	658	7
4	News Media	977	11	1
5	News Media	652	1312	22
6	News/Politics	600	1236	26
7	Indonesia	386	416	8
8	Foreign language (unknown)	495	312	5
9	Unclassified	372	589	25
10	Dera Sacha Sauda	1732	54	205
11	News about Russia	780	30	18
12	Shopping	1153	226	11
13	Greetings	476	1010	11
14	Greetings	439	1347	5
15	Science and animals	521	108	13
16	One Direction	388	1085	10
17	Foreign Language (Italian)	553	47	14
18	TV/Music	598	722	5
19	Music Videos	649	14	30
20	Nepal	393	748	125

After pruning out the foreign language communities in the dataset and some that were difficult to classify (Communities 0, 7-9, and 17), we can further group these onlookers into three broad interest groups: Casual Users (Communities 1-3, 12-14, 18, and 19), News and Pundits (4-6, and 13), and Engaged Users (20). We divided these subgroups based on the topics that were strongest in each, but these subgroups correlated with the number of quake-related tweets that each sent. They are described in the table below.

Category	Proportion	Quake-Related Tweets/ Week	Topic Affinities
Casual Onlookers	46%	3	Entertainment, greetings
News Onlookers	25%	6	News and politics
Engaged Onlookers	12%	10	Nepal

Casual Onlookers. Onlookers in these communities showed high activity but low engagement with the disaster, sending an average total of three quake-related tweets a week. Their primary topics of discussion were entertainment, or greetings and positive sentiments. This is the largest group.

News Onlookers. These accounts are either the accounts of professional news outlets or amateur pundits. We find low average quake-related tweets in this group as well: users sent an average of six relevant tweets per week. News outlets generally moved from one topic to another quickly, and pundits only sustained interest in the topics that appealed to them.

Engaged Onlookers. This group sent the most quake-related tweets of all users; the strongest LDA topics in this group were two "Nepal earthquake" topics. However, users in this community had few other topics in common with each other.

This breakdown suggests a model for predicting the number of tweets onlookers send about events. There will be roughly three categories of onlookers: Casual Onlookers, News Onlookers, and Engaged Onlookers. Casual Onlookers will consist of roughly 50% of onlookers, and will send only a few tweets over the first three weeks. Membership in this group is predicted by an interest in entertainment topics. The number of News Onlookers will be half the size of the Casual Onlookers, but they will be roughly twice as engaged. An onlookers's affinity to this group will be predicted by a general interest in news. Finally, the Engaged Onlookers will send 10-20 times as many tweets as the Casual Onlookers, and will comprise slightly over 10% of onlookers. This group sends the most tweets about an event, but membership in this group cannot be predicted from their preexisting interests.

Conclusion

We find that it is easy to predict shallow engagement with a disaster on Twitter, but difficult to anticipate sustained interest. Onlookers who tweet about entertainment are likely to pass on at least a few messages about donating money because entertainers are likely to post these messages, and fans are likely to retweet them. On the other hand, onlookers who tweet more about an event are likely to have preexisting interests that intersect with a particular aspect of the disaster, but the relevant interests are hard to predict because doing so would require knowing the nature of the disaster ahead of time. For example, to know the Hindu extremist would tweet about rumors of coerced conversions in Nepal, we would have to predict a crisis that would produce such rumors.

Additionally, we acknowledge that our model is derived from a single case study. As such, we treat it as provisional pending further experiments. We hope to confirm this model with future work.

Bibliography

Bastian, M., Heymann, S., Jacomy, M. et al. (2009). Gephi: an open source software for exploring and manipulating networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154./1009 (accessed 22 February 2016).

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**: 993–1022 (accessed 30 June 2014).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10). doi:10.1088/1742-5468/2008/10/P10008 (accessed 17 October 2013).

Bravo, C. A. and Hoffman-Goetz, L. (2015). Tweeting About Prostate and Testicular Cancers: What Are Individuals Saying in Their Discussions About the 2013 Movember Canada Campaign? *Journal of Cancer Education*. **1**(8). doi:10.1007/s13187-015-0838-8 (accessed 20 February 2016).

Burns, R. (2015). Digital Humanitarianism and the Geospatial Web: Emerging Modes of Mapping and the Transformation of Humanitarian Practices Thesis https://digital.lib.washington.edu:443/researchworks/handle/1773/33947 (accessed 20 February 2016).

Dugdale, J., Walle, B. Van de and Koeppinghoff, C. (2012). Social media and SMS in the haiti earthquake. ACM Press, pp. 713. doi:10.1145/2187980.2188189. http://dl.acm.org/citation.cfm?doid=2187980.2188189 (accessed 20 February 2016).

Gasso Climent, C. (2015). Twitter as a social marketing tool: modifying tweeting behavior in order to encourage donations. Info:eu-repo/semantics/bachelorThesis http://essay.utwente.nl/68039/ (accessed 20 February 2016).

Gladwell, M. (2011). *Outliers: The Story of Success*. Reprint edition. Back Bay Books.

Kaewkitipong, L., Chen, C. C. and Ractham, P. (2016). A community-based approach to sharing knowledge before, during, and after crisis events: A case study from Thailand. *Computers in Human Behavior*, **54**: 653–66 doi:10.1016/j.chb.2015.07.063 (accessed 20 February 2016).

Madianou, M. (2015). Digital Inequality and Second-Order Disasters: Social Media in the Typhoon Haiyan Recovery. *Social Media + Society*, **1**(2). doi:10.1177/2056305115603386 (accessed 20 February 2016).

Malcolm, K. (2016). How Social Media Affects the Annual Fund Revenues of Nonprofit Organizations. *Walden Dissertations and Doctoral Studies* http://scholarworks.waldenu.edu/dissertations/2005.

Mazaid, P. N. H. and A. D. and D. F. and M. H. and W. M. and M. (2011). Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?. http://ictlogy.net/bibliography/reports/projects.php?idp=2170 (accessed 15 May 2014).

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Amherst, MA: UMass Amherst http://mallet.cs.umass.edu.

Meng, Q., Zhang, N., Zhao, X., Li, F. and Guan, X. (2015). The governance strategies for public emergencies on social media and their effects: a case study based on the microblog data. *Electronic Markets*, **26**(1): 15–29 doi:10.1007/s12525-015-0202-1 (accessed 20 February 2016).

Morozov, E. (2012). *The Net Delusion: The Dark Side of Internet Freedom*. Reprint edition. New York: PublicAffairs.

Morozov, E. (2014). *To Save Everything, Click Here: The Folly of Technological Solutionism*. First Trade Paper Edition edition. New York: PublicAffairs.

Omilion-Hodges, L. M. and McClain, K. L. (2016). University use of social media and the crisis lifecycle: Organizational messages, first information responders' reactions, reframed messages and dissemination patterns. *Computers in Human Behavior*, **54**: 630–38 doi:10.1016/j.chb.2015.06.002 (accessed 20 February 2016).

Palen, L. (2008). Online social media in crisis events. *Educause Quarterly*, **31**(3): 76–78 (accessed 20 February 2016).

Platt, A., Hood, C. and Citrin, L. (2011). From Earthquakes to'# morecowbell': Identifying Sub-topics in Social Network Communications. *Privacy, Security, Risk and Trust (passat), 2011 Ieee Third International Conference on and 2011 Ieee Third International Conference on Social Computing (socialcom)*. IEEE, pp. 541–44 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6113164 (accessed 17 May 2014).

Pohl, D., Bouchachia, A. and Hellwagner, H. (2013a). Supporting Crisis Management via Detection of Sub-Events in Social Networks. *International Journal of Information Systems for*

Crisis Response and Management (IJISCRAM), **5**(3): 20–36 (accessed 17 May 2014).

Pohl, D., Bouchachia, A. and Hellwagner, H. (2013b). Social media for crisis management: clustering approaches for sub-event detection. *Multimedia Tools and Applications*, pp. 1–32 (accessed 17 May 2014).

Raine, L., Purcell, K. and Smith, A. (2016). The Social Side of the Internet | Pew Research Center *Pew Research Center: Numbers, Facts and Trends Shaping Your World* http://www.pewinternet.org/2011/01/18/the-social-side-of-the-internet/ (accessed 21 February 2016).

Regalado, R. V. J., McHale, K., Dela Cruz, B., Garcia, J. P. F., Ma, C., Kalaw, D. F. and Lu, V. E. (2015). FILIET: An Information Extraction System for Filipino Disaster-Related Tweets. Manila, Philippines: De la Salle University. http://www.dlsu.edu.ph/conferences/dlsu_research_congress/2015/proceedings/SEE/010-HCT_Regalado_RJ.pdf (accessed 20 February 2016).

Shirky, C. (2011). Political Power of Social Media - Technology, the Public Sphere Sphere, and Political Change, The. *Foreign Affairs*, **90**: 28.

Shulman, S. W. (2009). The Case Against Mass E-mails: Perverse Incentives and Low Quality Public Participation in U.S. Federal Rulemaking. *Policy & Internet*, **1**(1): 23–53 doi:10.2202/1944-2866.1010 (accessed 21 February 2016).

Tufekci, Z. and Wilson, C. (2012). Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square. *Journal of Communication*, **62**(2): 363–79 doi:10.1111/j.1460-2466.2012.01629.x (accessed 15 May 2014).

Xiao, Y., Huang, Q. and Wu, K. (2015). Understanding social media data for disaster management. *Natural Hazards*, **79**(3): 1663–79 doi:10.1007/s11069-015-1918-0 (accessed 20 February 2016).

Full Stack DH: Building a Virtual Research Environment on a Raspberry PI

James Dakin Smithies
james.smithies@protonmail.ch
King's College London, United Kingdom

Ian Hodder has recently pointed to a "return to things" in the humanities and social sciences, a mode of analysis that explores the relationships between people and the objects we use to construct and make sense of the world (Hodder, 2014, p.19). In digital humanities we can see this in Matthew Kirschenbaum's focus on the forensics of computer hard disks (Kirschenbaum, 2007), the development of platform studies to investigate the relationship between computing culture and the consoles and other hardware that enables it (Montfort and Bogost, 2009), and the appearance of maker cultures that seek to explore the humanities through practical experimentation (Dieter and Lovink, 2014). It suggests a desire to pay attention to and interact with the material world, rather than retreating to a purely digital one. Some commentators go beyond this. They propose that entanglement with material objects represents a ground of being for humans and their societies, that it presents a postphenomenological "dialectic of dependence and dependency between humans and things" worthy of deep contemplation (Hodder, 2014, p.19). People rely on the things they have created to such a degree, so the argument goes, that our identity has become inseparable from them. In Donald Ihde's original conception, it amounts to "recognition that "consciousness" is an abstraction, that experience in its deeper and broader sense entails its embeddedness in both the physical or material world and its cultural-social dimensions" (Ihde, 2009). Knowledge, art, religion, and science are entangled, in turn, with books, oil paint, churches, and laboratories: "thing theory" grounds epistemology in the myriad interactions between the physical and non-physical world (Preda, 1999).

If we apply these insights to digital infrastructure we begin to see how humanists have become entangled with complex systems, a situation that might prompt us to pause for thought. Analog books, archives, and libraries presuppose a degree of entanglement with the material world, undoubtedly, but those are relatively well understood: we have had centuries to understand and critique them. Digital infrastructure, however, is rarely "symbolically or politically construed" (Knorr-Cetina, 1992, p.119). It is merely requested in an email to a manager or Information Technology (IT) helpdesk, or held to be something humanists need to do their work. Little attempt is made to define the critical ground or, much less, to understand the object of enquiry prior to investment. Rather, in denial of the epistemological significance of things, humanities infrastructure is treated as something we should merely go to the store or work with our IT department to buy. The result has been an ongoing failure to provide the kind of infrastructure needed by humanities researchers, a misalignment of the debate, and often a rejection of the very notion of digital infrastructure itself.

As Matt Ratto points out, so-called 'critical making' provides new ways of dealing with difficult issues like this. Rather than attempting to solve complex problems in their entirety, critical making encourages the development of prototypes and basic models in the context of wider critical discourses, thus blending "practice-based engagement with pragmatic and theoretical issues" and fostering the possibility that creative solutions will be found to long-standing problems. He suggests the approach can be particularly useful in the context of so-called "wicked problems" (Ratto, 2011, p.253), defined by architectural scholar Horst Rittel in the 1960s and 1970s (Rittel and Webber, 1973). This class of problem is characterized by the existence of "many clients and decision makers with conflicting values, and where the ramifications in the whole system are thoroughly confusing" (Rittel was interested in problems associated with large-scale planning projects) (Churchman, 1967, p.B141). Significantly, he claimed there is a moral element to such problems, in that it is immoral to solve only one component of a wicked problem when such an approach will leave the larger issue unresolved. Prototyping and critical making can thus be positioned, not as inadequate tinkering, but as a mode of activity well-suited to the resolution of very complex problems. In this way we come to the intersection of critical making, cyber-infrastructure, and the humanities.

This project migrated my personal website jamessmithies.org from Wordpress.org (a free, fully hosted service) to a home server running on a Raspberry Pi 2 Model B minimal computer, a device built in the United Kingdom at Sony's manufacturing plant in Pencoed in South Wales and supported by a registered charity: the Raspberry Pi Foundation. The computer measures 85.60mm x 56mm x 21mm (or roughly 3.37″ x 2.21″ x 0.83″), has 1GB of Random Access Memory (RAM) and is powered by a 900MHz quad-core ARM Cortex-A7 Central Processing Unit (CPU). The VRE application is built using Django, a Python-based web framework designed for newspaper websites but now deployed in a wide variety of scenarios. The social media service Pinterest is one of the largest services to use it, with over 46 million unique visitors between 2011 and 2015 (Statista, 2015). The framework is thus highly adaptable, and could be used to develop almost any functionality a humanities researcher might need. The website is served by the Gunicorn application server and light-weight Nginx web server (used by NASA), with content saved in Postgres, one of the more advanced database systems available. All of these products are available free through the open source community. They require a reasonable level of technical proficiency to install and configure but there are many tutorials available online and their user communities share knowledge openly. It speaks to an interesting aspect of this project. Although there is a massive gap between jamessmithies.org and well-funded cyber-infrastructure projects, the nature of the open source software movement means there is only a small gap (if any) in terms of scalability and potential functionality.

One of the most powerful things about the project - in both technical and tactical terms - is the level of control conferred by the architecture of the 'stack'. Not only is the Pi itself accessible and configurable, but its operating system can be changed, and Gunicorn and Nginx can be configured at both an administrative level and through their core code base. Django can be programmed to support an extremely wide range of functionality. To extend the metaphor of control towards the incomprehensibly large infrastructures used by multi-national digital corporates (to escape the criticism that the Pi is a fundamentally limited device, or a mere toy), static files like CSS style sheets and images are hosted on the Amazon Web Services (AWS) cloud, integrating the Pi with a truly enormous global data infrastructure. These could have been hosted on the Pi, but it is considered best practice to deliver them separately for Django projects. It means, essentially, that much of the 'heavy-lifting' has been outsourced to a high performance computer, allowing almost limitless options for expansion of the site.

Perhaps counter-intuitively given the dominance of 'bigger is better' cyberinfrastructure discourse, the migration from Wordpress.org servers to a lowly Raspberry Pi has produced a personal VRE capable of significant further development. The intention is not necessarily to create a finished and reproducible product, but to take control of – and experiment with - all aspects of the computing architecture in order to gain a better understanding of my scholarly infrastructure needs, from the hardware the site runs on, to maintenance of the Internet domain name, the content management system that helps me organize content, and the firewalls that secure it from malicious actors. The conclusion after this phase of the project is that issues like ethical hardware, net neutrality, data sovereignty and security, and the ability to extend and configure the code that supports my research activities, are central to my work – and identity - as a humanities scholar.

Bibliography

Anon., (2015). Pinterest: Unique U.S. Visitors 2015. *Statista*. April 2015. http://www.statista.com/statistics/277694/number-of-unique-us-visitors-to-pinterestcom/ (accessed July 11, 2015).

Churchman, C.W. (1967). Guest Editorial: Wicked Problems. *Management Science*, **4**: 141-42.

Dieter, M. and Lovink, G. (2014). Theses on Making in the Digital Age. In *Critical Making*. California: Garnet Hertz.

Hodder, I. (2014). The Entanglements of Humans and Things: A Long-Term View. *New Literary History*, **45**: 19-36.

Ihde, D. (2009). *Postphenomenology and Technoscience: The Peking University Lectures.* Albany: SUNY Press.

Kirschenbaum, M. (2007). *Mechanisms: New Media and the-Forensic Imagination.* Cambridge Mass.: The MIT Press.

Knorr-Cetina, K. (1992). The Couch, the Cathedral, and the Laboratory: On the Relationship between Experiment and Laboratory in Science. In Pickering, A. (ed.), *Science as Practice and Culture*, ed. Chicago: University of Chicago Press, pp. 113-38.

Montfort, N. and Bogost, I. (2009). *Racing the Beam: The Atari Video Computer System*. Cambridge Mass.: The MIT Press.

Preda, A. (1999). The Turn to Things: Arguments for a Sociological Theory of Things. *The Sociological Quarterly*, **2**: 347–66.

Ratto, M. (2011). Critical Making: Conceptual and Material Studies in Technology and Social Life. *The Information Society*, **4**: 252-60.

Rittel, H.W.J. and Webber, M.M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, **2**: 155-69.

Sayers, J. (2014). The Relevance of Remaking. *Maker Lab in the Humanities*. November 24, 2014. http://maker.uvic.ca/remaking/ (accessed 01 March 2016).

SpotiBot-Turing testing Spotify

Pelle Snickars
pelle.snickars@umu.se
Umea university, Sweden

Producing and coding bot 'listeners' has today become almost as easy as automated music production has been for years. Machines can thus both 'create'—and 'listen' to 'music' (whatever we mean by these categories). In fact, such notions are capricious within the contemporary streaming music landscape. The project, "Streaming Heritage: Following Files in Digital Music Distribution" (financed by the Swedish Research Council) studies emerging streaming media cultures in general, and the music service Spotify in particular, with a bearing on the digital challenges posed by direct access to musical heritage. Rediscovering older music is key for Spotify, and the project is hence on the one hand geared towards investigating the institutional challenges of streaming musical heritage, and on the other hand—and foremost—to develop new digital research methods. Situated at HUMlab (Umeå university) part of the project is essentially about Turing testing Spotify. Building on the tradition of 'breaching experiments' in ethnomethodology, my research group seeks to break into the hidden infrastructures of digital music distribution in order to study its underlying norms and structures. The key idea is to follow files' (rather than the people making or using them) on their distributive journey through the streaming ecosystem. The setting include the distribution and aggregation of self-produced music/sounds through Spotify; the set-up of our own record label (for research purposes); the programming of bots to inform, explore, mimic, and ultimately subvert notions of usage and listening; the tracing of Spotify's history through constantly changing interfaces (web archiving and documenting these). Research questions range from various way how streaming music is commodified? What sounds are perceived as music (or not) according to Spotify and adjacent aggregating services? How is metadata generated, ordered, and valued—and what kind of metadata is actually available? What normative world views are promoted and materialized by streaming architectures? What kind of infrastructures proliferate behind the surfaces of on-demand services?

My presentation departs from the fact that one-fifth of Spotify's catalogue of 30 million songs haven't once been listened to at all. Under the computational hood of streaming services all streams are equal, and every stream thus means (potentially) increased revenue from advertisers. Spotify is hence likely to include—rather than reject—various forms of (semi-)automated music, sounds and (audio)bots. At HUMlab we therefore set up an experiment—SpotiBot—with the purpose to determine if it was possible to provoke, or even to some extent undermine, the Spotify business model (based on the 30 second royalty rule). Royalties from Spotify are only disbursed once a song is registered as a play, which happens after 30 seconds. The SpotiBot engine was be used to play a single track repeatedly (both self-produced music and Abba's ”Dancing Queen"), during less and more than 30 seconds, and with a fixed repetition scheme running from 10 to *n* times, simultaneously with different Spotify account. Based on a set of tools provided by Selenium the SpotiBot engine automated the Spotify web client by simulating user interaction within the web interface. From a computational perspective the Spotify web client appeared as black box; the logics that the Spotify application was governed by was, for example, not known in advance, and the web page structure (in HTML) and client side scripting complex. It was not doable within the short experiment to gain a fuller understanding of the dialogue between the client and the server. As a consequence, the development of the SpotiBot-experiment was (to some extent) based on 'trial and error' how the client behaved, and what kind of data was sent from the server for different user actions. Using a single virtual machine—hidden behind only one proxy IP—the results nevertheless indicate that it is possible to automatically play tracks for thousands of repetitions that exceeds the royalty rule. Even if we encountered a number of problems and deviations that interrupted the client execution, the Spotify business model can be tampered with. In other words, one might ask what happens when—not if—streaming bots approximate human listener behavior in such a way that it becomes impossible to distinguish between a human and a machine? Streaming fraud, as it has been labeled, then runs the risk of undermining the economic revenue models of streaming services as Spotify.

Reconstruction of Labour Relations in the North Sea Region in the Late Middle Ages: Spatio-Temporal Analysis Using Historical GIS, Taxation Sources, and Coin Finds

Rombert Stapel
rombert.stapel@iisg.nl
International Institute of Social History

Introduction

In this paper a range of maps, documents, data and archaeological finds are brought together in a historical GIS (Geographical Information System) to reconstruct different forms of labour and especially labour relations in existence in late medieval Northwest Europe. Starting from the Late Middle Ages, England and the Low Countries laid the groundwork for becoming two of the dominant powers in Europe and the world. The question is how this process of growth began. Often it has been coined with the rise of capitalism and the rise of wage labour as the predominant form of labour (e.g. van Bavel, 2007). The issue also touches on the Great Divergence debate. This debate centres around why and at what point in time Europe became the dominant continent in the world, surpassing the power and wealth of China, India, Japan and the Ottoman Empire (Pomeranz, 2000). England and the Low Countries play an important role in this debate and quite often the level of (real) wages is used as a determinant to study the emerging differences between Europe and other parts of the world (Allen, 2001). This opens up the question, however, how many people in a certain time and place actually worked for wages: an issue which questions the validity of real wages as (the only) measured variable (Lucassen, 2016).

Labour relations

Labour relations can be defined as "for or with whom one works and under what rules" (Hofmeester et al., 2015). In the "Global Collaboratory on the History of Labour Relations, 1500-2000", a collaboration of researchers from all over the world, study these labour relations and especially shifts between them. To effectively compare different parts of the world across time, a uniform way of entering and presenting data as well as a taxonomy has been created (Figure 1). One of the advantages of this taxonomy is that it encompasses the entire population of an area, working and non-working, forces researchers to think about female labour participation (even if it is omitted in their sources), and includes various forms of labour, from reciprocal labour to wage labour, self-employment or slavery.

Studying societies other than modern ones can be especially difficult however. Lacking anything like modern census data, one has to be more creative to study different forms of labour across time and space and bring together a wide range of data and information. One may think about information on land use, vicinity of roads, waterways or coastlines, all proving modes of transportation, the vicinity of towns and monasteries, demographic density, documents related to taxation and many other. GIS is best equipped to present such combinations of different data, and is therefore at the core of the following two areas for which the labour relations are reconstructed.

Reconstruction England/Wales

To come to a reconstruction of the labour relations in England and Wales a range of sources are combined, including the 1378-1381 Poll Tax records (Fenwick, 1998), the muster rolls of 1522, information on medieval markets (Keene et al., 2002; Keene and Letters, 2004), and GIS shapefiles of historical parish boundaries (Southall and Burton, 2004). From these sources regional variation in demography, market presence, and occupational structure can be extracted – albeit with many caveats. One of the main sources used for the reconstruction, however, are the tens of thousands of archaeological coin finds, by both amateurs and professionals. These are made available through the Portable Antiquities Scheme (PAS) website (The British Museum, 2013-2015) (Figure 2).

The principal assumption behind using the coin finds is that the presence of small denominations, valued at a day's wage or less, can function as a determinant of wage labour (Lucassen, 2014). Used in combination with figures on mint output, the coin finds can therefore be used to study the relative presence of wage labour; as well as regional differences and developments over time.

As the PAS database was formed by different people over a long period of time, various issues caused by inconsistent or erroneous data entry had to be solved. Much effort has gone into cleaning (especially place names) in the Portable Antiquities Scheme database and supplementing the numerous missing geographical coordinates. The following step is linking the coin finds to the data mentioned earlier, a complicated process that has started but for which much still has to be done in the coming months. In the end, the combination of available data provides us with an extensive toolbox to reconstruct the labour relations in late medieval England and Wales and to study the mechanisms that influence these labour relations, causing regional and temporal variation.

Reconstruction Low Countries

For the reconstruction of the labour relations on the other side of the North Sea, in the Low Countries a different approach has been chosen, although here too coin finds may be used in the end (De Nederlandsche Bank, 1997-2015). Starting point is the County of Holland. In 1494 and 1514 two sets of questionnaires were produced by the Burgundian/Habsburg rulers, intended to allocate a new round of taxes. The questionnaires had to be answered by representatives of all towns and heerlijkheden (a feudal administrative-judicial unit, precursor of modern municipalities) in Holland. They asked for information about the number of dwellings and parishioners; economic activities; how these economic activities developed in the past couple of years; land ownership; and tax-related issues. The questionnaires have been used to assess the state of the economy in Holland and study per capita growth (van Zanden, 2002).

Never before have these questionnaires been mapped, which, for instance, would allow to combine the information with land use, the vicinity of monasteries and modes of transportation, and study in more detail regional variance in the county and – because of the nature of the questionnaires – changes over time between 1477 and 1514. For this purpose, a historical GIS of the administrative-judicial boundaries in this and some neighbouring counties was created from scratch. A wide range of digitised historical maps from the sixteenth to eighteenth century were used, in combination to the historical atlas of the Netherlands drawn by Anton Beekman at the beginning of the twentieth century (Figure 3; results in Figure 4). The maps were georeferenced and combined with information on natural and/or current (sometimes unchanged) municipality borders to reconstruct the location of the historical boundaries.

One of the immediate advantages was that using this GIS map, silent omissions in the questionnaires became visible (e.g. certain heerlijkheden that were not mentioned explicitly in the sources). Moreover, the advantage of using the historical boundaries, instead of just looking at the location of the villages and creating Voronoi diagrams for instance also becomes clear, as the maps provide much more detail and display patterns more clearly: see Figure 5.

By combining the questionnaires and the newly created geographical information with assumptions on household size, child and female labour participation, information on religious houses (using Goudriaan and Stuyvenberg, 2008), and other information a tentative reconstruction is made on the full range of labour relations present in the County of Holland in the Late Middle Ages.

Conclusion

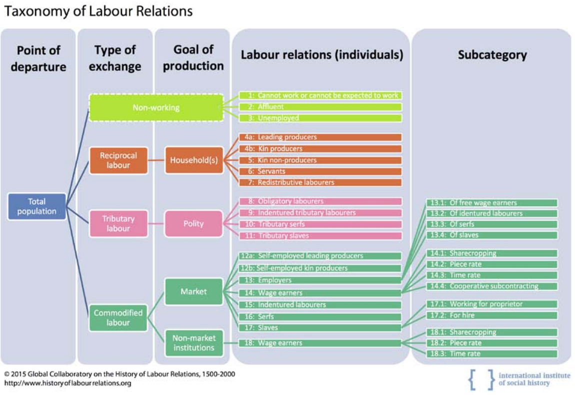


Figure 1. Taxonomy of the Global Collaboratory on the History of Labour Relations, 1500-2000.

In the end, the goal is to create a map that shows the presence and absence of various forms of labour and labour relations in the North Sea region. This allows us to study this part of the world, just before two new world powers emerged from here, and the role labour plays in this development. By using GIS (instead of displaying the information in tables for instance), the full potential offered by geographical information systems is made available: e.g. easily combining different forms of information, pinpointing developments to a certain point in time and space. Large-scale regional variation can therefore become a point of study, instead of hindering or difficult to grasp principle.

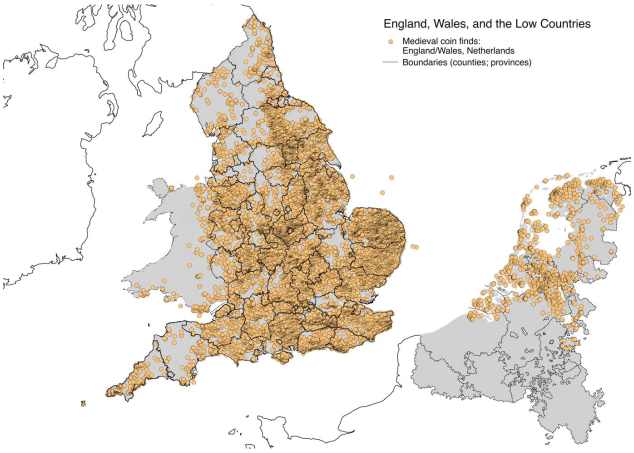

England, Wales, and the Low Countries
Medieval coin finds: England/Wales, Netherlands
Boundaries (counties; provinces)

Figure 2. Archaeological coin finds in England/Wales and the Netherlands (Northern Low Countries). Note that at the moment the finds in the Netherlands and England/Wales cannot be compared easily due to national differences in registration.

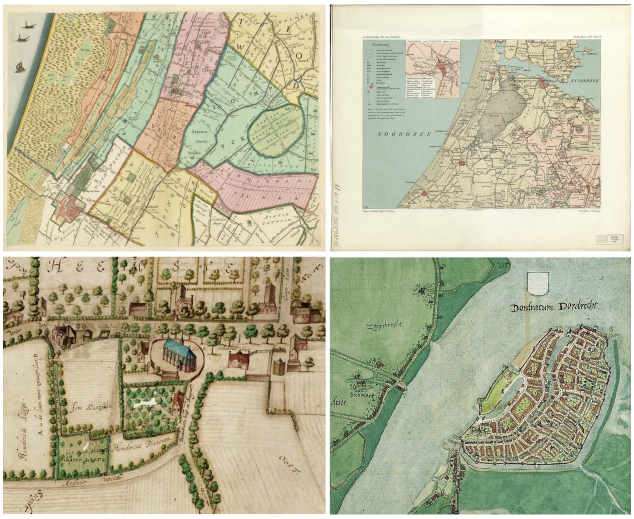


Figure 3. Excerpt of geographical sources: various 16th-18th century maps, historical atlases.

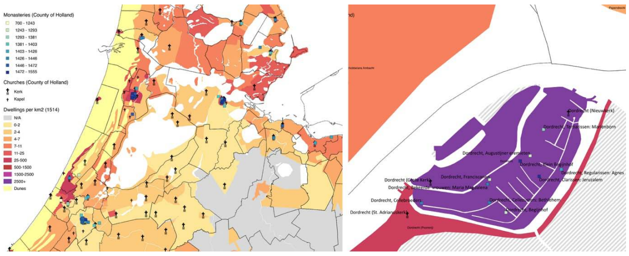


Figure 4. Result: Historical GIS of administrative-judicial units in Low Countries around 1500 (with churches; monasteries; present-day municipality boundaries).

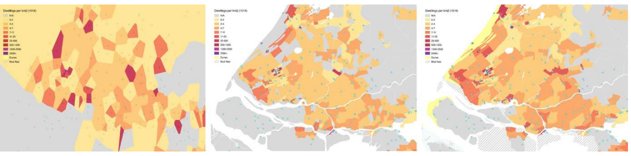


Figure 5. Thought process and development of mapping medieval tax/population records in Holland: from Voronoi diagrams surrounding the towns and villages mentioned in the sources; to mapping these towns and villages; and finally locating and assessing the areas not explicitly mentioned in the sources.

Bibliography

Allen, R. C. (2001). The Great Divergence in European Wages and Prices from the Middle Ages to the First World War. *Explorations in Economic History*, *38*: 411–447.

De Nederlandsche Bank (1997-2015). *NUMIS Muntvondsten Database*. URL http://www.dnb.nl/over-dnb/nationale-numismatische-collectie/numis/index.jsp (accessed 10.31.15).

Fenwick, C.C. (Ed.) (1998). *The poll taxes of 1377, 1379, and 1381, Records of social and economic history*. Published for the Bristish Academy by Oxford University Press, Oxford; New York.

Goudriaan, K., Stuyvenberg, B. (2008). *Kloosterlijst: beknopt overzicht van de Nederlandse kloosters in de Middeleeuwen*.

Hofmeester, K., Lucassen, J.M.W.G., Lucassen, L.A.C.J., Stapel, R.J., Zijdeman, R. (2015). *The Global Collaboratory on the History of Labour Relations, 1500-2000: Background, Set-Up, Taxonomy, and Applications*. URL http://www.historyoflabourrelations.org/ (accessed 10.31.15)

Keene, D., Galloway, J., Murphy, M., Myhill, O. (2002). *Metropolitan Market Networks, c. 1300-1600; London, its Region and the Economy of England*.

Keene, D., Letters, S. (2004). *Markets and Fairs in England and Wales to 1516*.

Lucassen, J.M.W.G. (2014). Deep Monetisation: The Case of the Netherlands 1200-1940. *TSEG*, *11*: 73–122.

Lucassen, L.A.C.J. (2016, forthcoming). Working Together. New Directions in Global Labour History. *Journal of Global History*, *11*(1).

Pomeranz, K. (2000). *The Great Divergence: China, Europe, and the Making of the Modern World Economy*. Princeton University Press, Princeton, N.J.

Southall, H.R., Burton, N. (2004). *GIS of the Ancient Parishes of England and Wales, 1500-1850*.

The British Museum (2013-2015). *Portable Antiquities Scheme*. URL http://finds.org.uk/ (accessed 10.31.15).

Van Bavel, B. (2007). The Transition in the Low Countries: Wage Labour as an Indicator of the Rise of Capitalism in the Countryside, 1300-1700. *Past & Present*, *195*: 286–303.

Van Zanden, J.L. (2002). Taking the measure of the early modern economy: Historical national accounts for Holland in 1510/14. *European Review of Economic History*, **6**: 131–163.

Curating Just-In-Time Datasets from the Web

Todd Suomela
todd.suomela@ualberta.ca
University of Alberta, Canada

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca
University of Alberta, Canada

Ryan Chartier
rchartier@ualberta.ca
University of Alberta, Canada

The Web is now deeply integrated into contemporary culture, and scholars interested in current phenomenon cannot afford to ignore it, however, collecting data from the web is not easy. The web is based on a mix of continually changing technical standards which make creating an archival copy of a web site for scholarly reference very difficult. Without such a copy there is no way for future scholars to question the interpretations we make today or reinterpret the phenomenon in light of new evidence tomorrow.

Researchers and organizations, such as the Internet Archive, are attempting to preserve portions of the web for future retrieval, but much of the web disappears quickly. A 2014 study of web links in scholarly papers found that 1 out 5 scholarly papers contains links to web URLs which no longer function or no longer exist (Klein et al., 2014). The need for humanists to recognize the value of web archives to historical research is especially acute. Researchers cannot engage recent cultural histories and ignore the culture of the web (Milligan, 2012).

The challenge of web archiving is especially acute for rapidly changing stories which track specific events. Discussions about controversial topics, such as GamerGate,[1] take place across multiple websites, use multiple forms of media, and occur in very different discourse communities. Underneath the different social worlds gathered around online communities there is an incredibly diverse set of technological platforms which require customized strategies for tracking and collecting data. This paper will:

- Describe the key challenges for researchers wishing to collect just-in-time archives of web based cultural phenomenon.
- Put current challenges in an historical context of differing goals between web developers and web archivists.
- Propose some social and technical solutions to improve the situation, and
- Introduce a set of tools to help researchers engaged in these areas.

Challenges

Dynamic changes in online content present one of the unique challenges for gathering contemporary web discourse. Most internet users are familiar with the constantly updating nature of Twitter and other social media platforms. Social media platforms present a challenge for web archivists because of their technological structure and commercial ownership. The speed of updates on social media requires specialized tools to download, especially in large quantities. A researcher needs deep technological knowledge of these tools and the application programming interface (API) provided by the website in order to build a reliable and useful corpus. On the legal side, the terms of service affect the types of information that can be gathered by researchers and how that data can be analyzed or shared with other researchers.

Other commercial sites, such as news media web sites, often host comment threads where internet users can post their opinions on the topics covered in the main story. It is relatively trivial to download the main content of a news story posted on the web but collecting the comment stream may present a challenge because the comments may be hosted by another website service or may be displayed dynamically as a user scrolls further through a web page. In such cases the default web archiving tools may not be sufficient. Web discussion forums present yet another technical challenge.

Researchers often frame their questions about web phenomena by describing a topic that they wish to study. But the architecture of the web is built around the key idea of a web site, a particular set of files which may include many different types of media including text, images, and video, and is hosted by a particular business, institution, or individual. These web sites are identified by the Uniform Resource Locators people type into their browsers in order to navigate to a web page.

The tools used to archive the web are built on this technical background for dealing with URLs, APIs, REST, RSS, and other interfaces which human beings do not usually interact with. In the language of web archive software the unit of research is the seed, or base URL, from which data can be harvested. For the researcher the unit of work is the topic. Negotiating between these two conceptions of how online research should work is a major social challenge for any type of internet research. Researchers and web users just want to see the content, but automating the collection of that content means reproducing a complicated software experience which has gradually been built by web developers and web browsers over the past 25 years.

Humanities researchers have traditionally relied on stable or slowly changing content. Efforts by humanities scholars have been made to adapt to the changes in digital content represented by the web. Some universities have set up web labs for collecting and analyzing web data. One key task of these labs has been building subcollections from the overall web in order to further the study of particular topics (Arms et al., 2009). One of the key insights from our work is the need to continue building strong collaborations between multiple fields. Libraries and the Internet Archive need input from digital humanists in order to understand their research questions and digital humanists need to understand the technological challenges of web archiving in order to collectively design systems which will help future researchers. The web, however, is constantly changing at multiple levels, ranging from the technology used to deliver content, the processes of creating content, commenting on content, and the distribution of information. Archiving the web for humanities research calls for changing the conceptual image of stable sources, collaborating with new communities, and adopting new technologies.

Solutions

The implication of the technological treadmill described above is that it becomes more and more difficult for a single researcher to adequately collect the web. There are two potential solutions to this problem: technological and social.

Computer scientists are working to build better web archive software which can integrate with social media in order to reduce the amount of administrative overhead needed to collect information on particular topics.[2] These tools will automate the selection of web sites to be archived, removing some of the human intervention needed to curate web materials. But simplifying the data gathering process today may make future explanations of the context of a collection more technical. For now researchers are dependent upon a mix of tools, often customized for specific uses, and mixing open source and commercial software.

In our research project we used a combination of open source tools, subscription services, and customized API calls. For gathering data from Twitter we used a program called twarc.[3] Customizations were made to improve the performance of the tool for our uses, which was tracking specific hashtags. The Twitter scraper was initially installed on a laptop belonging to a member of the research team, but when the number of tweets became too large for a laptop the program was moved to a cloud server provided by Compute Canada. The data from Twitter was stored in JSON and then transformed using standard libraries into files which could be analyzed for most frequent Twitter posters, most frequent URLs, and most frequent hashtags. Data from web pages was collected using the Archive-IT subscription service[4] provided by the Internet Archive and the wget[5] command line tool. Some specific websites, such as 4chan and 8chan, required the development of custom API interfaces to download material from relevant chat boards. Additional programs to download comments from

YouTube are currently being tested. We plan to document our recipes for using these different tools on methodi.ca,[6] the methods commons for text analysis.

Archiving the web involves many different institutions and disciplines. The largest players are the Internet Archive and various national libraries; the Internet Archive operates as a non-profit and has the most comprehensive collection of digital materials from the web. Unfortunately, the Internet Archive collections are not primarily built for researcher access and can be especially difficult to work with if you are investigating topics which cover multiple URLs or lengthy time periods. Any research project using their collections requires significant human labor. Libraries and museums can step in to fill some of the gaps by using services such as Archive-IT, which provides more curatorial control over the collection development process and also has a more robust search interface. In order to improve these tools, humanists will need to build connections with other disciplines, such as information and library science, computer science, and archival studies. Only by working together and extending our disciplinary horizons can we build the collections which current and future digital humanists can use to study our current era.

One final social issue of importance are the legal and ethical implications of gathering large amounts of data from the web. We will not discuss these issues in great depth in this paper but they do need to be acknowledged because they constrain some of the actions which can be taken in web data gathering. In our project on GamerGate we have looked closely at the ethical implications of sharing data gathered from social media. The dataset we shared online includes an appendix on ethical issues related to data sharing.[7]

Bibliography

Arms, W.Y., Calimlim, M. and Walle, L. (2009). EScience in Practice: Lessons from the Cornell Web Lab. *D-Lib Magazine*, **15**(5/6). doi:10.1045/may2009-arms

Hathaway, J. (2014). What is GamerGate, and Why? An Explainer for Non-Geeks. *Gawker*. http://gawker.com/what-is-gamergate-and-why-an-explainer-for-non-geeks-1642909080

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K. and Tobin, R. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE*, **9**(12): doi:10.1371/journal.pone.0115253

Milligan, I. (2012). Mining the "Internet Graveyard": Rethinking the Historians' Toolkit. *Journal of the Canadian Historical Association*, **23**(2): 21-64. doi:10.7202/1015788ar

Notes

[1] This paper is the result of a larger project investigating the discourse surrounding GamerGate, an internet controversy about feminism and gaming, which grew dramatically in 2014. The paper presents some of the methods used by our research group to study GamerGate. For a brief non-academic explanation of GamerGate see Hathaway, 2014

[2] Some of these research groups are located at the Center for the Study of Digital Libraries at Texas A and M; Web Science and Digital Libraries Research Group at Old Dominion; and the Digital Library Research Laboratory at Virginia Tech.

[3] Github repository at https://github.com/edsu/twarc

[4] Web site https://archive-it.org/

[5] Web site https://www.gnu.org/software/wget/

[6] Web site http://methodi.ca/

[7] Rockwell, G., Suomela, T., 2015, "Gamergate Reactions", http://dx.doi.org/10.7939/DVN/10253 V5 [Version]

Read, Play, Build: Teaching Sherlock Holmes through Digital Humanities

Joanna Elizabeth Swafford
swafforj@newpaltz.edu
SUNY New Paltz, United States of America

The nineteenth century provides a perfect setting for a digital humanities class as a result of the similarities between the industrial and digital revolutions and between the proliferation of print and periodicals and the rise of blogs and Twitter. Arthur Conan Doyle's Sherlock Holmes stories are likewise the perfect subject matter: most of Holmes's cases revolve around the technology of the day, from cabinet photographs in "A Scandal in Bohemia" to the typewriter in "A Case of Identity." These connections enable students to learn about an earlier time period and literature while also historicizing their own technological moment. The resurgence in popularity of Holmes adaptations in recent years—*Sherlock* (BBC), *Elementary*, and *Mr. Holmes*, to name just a few—emphasizes these connections and also brings students into the classroom.

"Digital Tools for the 21st Century: Sherlock Holmes's London" (taught from the Fall of 2014 through the Fall of 2015) is an introductory digital humanities class that uses Holmes stories as a corpus on which to practice a wide variety of basic digital humanities methodologies and tools, including visualizations, digital archives and editions, mapping, and distant reading. "Digital Tools" unites theory and practice with a tripartite structure for each unit, which I have dubbed "Read, Play, Build." First, students read articles from books and blog posts about the pros and cons of each methodological approach. They then examine current projects to discuss the ways that each enhances scholarly fields and poses new research questions. Each unit concludes with an in-class lab component, in which they build a small project using a well-known tool. For example, in the archives and editions unit, students examine Jerome McGann's "Radiant Textuality," discuss the importance of preservation, access, and challenging the canon, examine *The Rossetti Archive* (and many others), and use the tool *Juxta Editions* to create their own, fully transcribed digital edition of a Sherlock Holmes story of their choice, using page images of the original printing in The Strand Magazine to learn about XML, basic bibliography, and best editorial practices. Likewise, in the unit on distant reading and topic modeling, they read Franco Moretti's "The Slaughterhouse of Literature" and Ted Underwood's blog posts on topic modeling, examine *Mining the Dispatch* to see that methodology in action, and topic model all 56 Sherlock Holmes short stories and analyze the trends in a blog post.

Although this class uses Holmes stories as base texts, it also situates these stories in their historical and cultural context by examining Victorian digital humanities projects from fields other than English. Students explore the *Proceedings of the Old Bailey*, a searchable archive of all court records in the Old Bailey from 1674-1913, to learn about crime, class, and the legal system. They also examine the *David Livingstone 1871 Field Diary* to learn about empire and Charles Booth's poverty maps from 1898-1899 from the *Charles Booth Online Archive* to learn about socioeconomic conditions in London and to compare Conan Doyle's fictional London to the actual city.

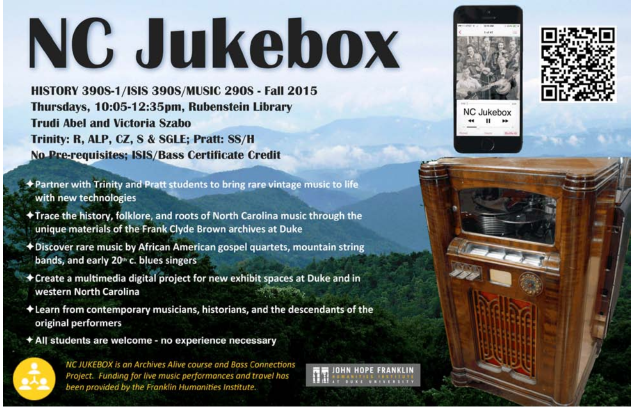
This paper will present the class in greater depth and will provide examples of the digital projects students collaboratively created—from contributing to the marginalia project *Book Traces* to making digital narrative maps of Holmes stories with Mapbox—in order to provide new models for student scholarship and their role in the future of the English departments and the humanities. It will also discuss the importance of the Holmes stories as the corpus for the class, as the character of Holmes himself provides a useful model for a digital humanist, especially when students may be unused to thinking about data in a humanities context, as he combines data science with humanities skills of close reading, archiving, and a love of literature and music.

Teaching Holmes with digital tools lets students build on the traditional humanities skills of close reading, understanding patterns, and using archives, and augments that scholarly toolkit by guiding students to a better understanding of rhetorical patterns and spatial significance, while also teaching them about collaboration, interdisciplinarity, and public humanities. In accordance with its public humanities focus, the course's materials, including the syllabus and assignments, are publicly available on the class's website (https://hawksites.newpaltz.edu/dhm293/). By melding research with project-based learning, this course enables students to engage with research and Victorian history more actively than is common at the introductory level.

The North Carolina Jukebox Project: Archives Alive and the Making of Digital Cultural Heritage

Victoria Szabo
ves4@duke.edu
Duke University, United States of America

The North Carolina (USA) Jukebox project transforms an inaccessible audio archive from the 1930s, of historic North Carolina folk music collected by Frank Clyde Brown, into a vital, publicly accessible digital archive and museum exhibition. Led by Trudi Abel, a librarian in the Rubenstein Special Collections Library at Duke University, and Victoria Szabo, a faculty member in Visual and Media Studies and Information Science + Studies at Duke, this interdisciplinary, collaborative effort also involves scholars in music and folklore, music and preservation librarians, digital media specialists, descendants of the original performers, and contemporary musicians who play this music professionally today (Archives Alive Initiative | Trinity College of Arts and` Sciences, 2016).



NC Jukebox

HISTORY 390S-1/ISIS 390S/MUSIC 290S - Fall 2015
Thursdays, 10:05-12:35pm, Rubenstein Library
Trudi Abel and Victoria Szabo
Trinity: R, ALP, CZ, S & SGLE; Pratt: SS/H
No Pre-requisites; ISIS/Bass Certificate Credit

✦Partner with Trinity and Pratt students to bring rare vintage music to life with new technologies
✦Trace the history, folklore, and roots of North Carolina music through the unique materials of the Frank Clyde Brown archives at Duke
✦Discover rare music by African American gospel quartets, mountain string bands, and early 20th c. blues singers
✦Create a multimedia digital project for new exhibit spaces at Duke and in western North Carolina
✦Learn from contemporary musicians, historians, and the descendants of the original performers
✦All students are welcome - no experience necessary

NC JUKEBOX is an Archives Alive course and Bass Connections Project. Funding for live music performances and travel has been provided by the Franklin Humanities Institute.

JOHN HOPE FRANKLIN HUMANITIES INSTITUTE AT DUKE UNIVERSITY

Figure 1. NC Jukebox Project Advertisement

As a teaching experience associated with the Library's Archives Alive initiative, the project offers opportunities for students from arts, music, computer science, and engineering programs to learn about the collection and develop an exhibition from initial concept to execution, and to do so in collaboration with a diverse set of mentors and collaborators who help them understand the histories and technologies involved, as well as stakes of their presentation choices. As a ongoing archival project, it demonstrates the challenges and opportunities inherent in working out a major library archive and preservation effort alongside a live curricular intervention and planned public exhibition. As a research project, it offers scholars in media studies a firsthand view of how material recording and playback technologies and their affordances help shape subsequent cultural histories, and affect what we can recirculate and share today. Taken together, these strands demonstrate that introducing digital cultural heritage project development as a shared objective enriches student learning, encourages library archiving and preservation projects to consider their public facing dimensions as they construct new resources, and offers digital humanities and media studies scholars meaningful opportunities to collaborate with colleagues in historically minded disciplines around new forms of scholarly production – in this case data-driven exhibitions at the Mountain Music Museum in Western NC, at the Rubenstein Special Collections Library at Duke, and online.

Our project begins with the songcatcher himself. In the 1930s Frank Clyde Brown, Duke Professor of English, and co-organizer of the North Carolina Folklore Society (1913) as Zeke Graves in our Library tells us, began recording and archiving Western North Carolina folk music (Graves, 2015). Following in the tradition of folklorist Alan Lomax, and songcatcher/musician Bascom Lunsford as chronicled by Loyal Jones, along with other famous songcatchers of the period, he drove around region capturing a range of singers and songs using the technology available in the period, a notebook and Dictaphone equipped with first wax cylinders and later aluminum cylinders (Jones and Forbes, 1984). Like us, Brown involved his students in the project as well, encouraging them to capture songs and research their origins. Today most of those recordings are still housed on wax cylinders and glass disks in the Duke Libraries, and in the Library of Congress, with about 400 songs having already been converted to digital formats. The rest are being converted as part of a substantial Council of Library and Information Resources grant.

In addition to learning about Brown, histories of the music, songcatching, and folklore practices of the period, students in a Fall 2015 NC Jukebox course began to work closely with the 400 digitized recordings we currently have available, developing metadata, transcribing songs, and organizing their materials in spreadsheet, blog, and database form. Our project also explores biographies of the singers, transcribes the songs as heard on the tapes in in comparison to other versions, and traces the Scotch-English history and contemporary analogues of the songs themselves through research in Child's *Ballads* and other key sources (Child et al., 2001). In addition, we have begun to demonstrate change over time and space through maps, patterns, flows, timelines, and networks of the music - a kind of distant listening, or viewing of its collection and presentation, with more to come. In the physical exhibits, interactive touchscreens, period photos, and hybrid analog-digital audio playback machines – a radio, a Jukebox, and perhaps a 78-playing phonograph – will invoke the historical conditions of production and reception of the music for diverse audiences.

As an historiographical research project, NC Jukebox is also offering opportunities to explore firsthand how social and material conditions affect the writing of cultural history itself. Over the course of this project we have learned about the history of songcatching and folklore as social and academic practices designed to verify expectations about a specific kind of musical heritage. Brown died before his work could be compiled into the published versions of his work. As his papers reveal, not all of the songs he collected were included in the final, posthumous collection of his work (Guide to the Frank Clyde Brown Papers, 1912-1974, 2016). His subsequent editors, like other before them who were seeking a pure musical tradition descended from that of the Scots-Irish settlers in the region, as seen in Ritchie and Orr's *Wayfaring Strangers,* for example, picked and chose songs to include in the published work (Ritchie, Orr and Parton, 2014). Their criteria are (helpfully for future researchers) sketched out in their editorial notes, which are also in the archives. Songs that were too popular, had been published, were too religious, or, perhaps most significantly, were from African American traditions, were excluded from the published collection, even if familiar from other sourdes like the Library of Congress Checklist of Recorded Songs (U.S Library of Congress, 1942). On another note, we also experienced the lingering effects of musical and cultural segregation first-hand as a class when our well-intentioned musical guest, a contemporary folksinger who is helping keep the Mountain Music traditions alive today, explained that he was going to perform a song as he had heard his "colored" neighbors sing it growing up (McKinney, 2015). This moment became part of a class conversation about representing tradition while at the same time acknowledging our contemporary perspectives upon it in how we frame the music.

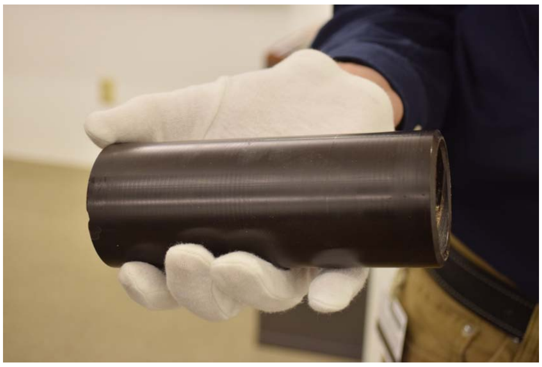


Figure 2. Wax Cylinder from the FCB collection

This project is also about the history of recording and reproduction technologies, which has deeply affected the content. Encountering recordings made from wax cylinders and glass disks, which included a white-glove visit from Special Collections as well as examining photographs of Brown and other "in the field," encouraged conversations around our continuously evolving standards and expectations for archiving and reproduction. Our students had to consider whether it was more "authentic" to leave in the hisses and crackles that had made it into the third generation audio files they were listening to, or whether they could and should attempt to clean up the sound quality so it was closer to the "original" source – the singers themselves. Were we archiving the archives, or the ur-performances? Our students also learned about and from Charles Bond, the onetime Duke undergraduate (and now lawyer in San Francisco) who serendipitously took it upon himself in the 1980s to transfer some of the existing recordings to reel-to-reel tapes, using a moog synthesizer to clean up some pops and crackles.

Further, we discovered how the limits of the recording medium itself reveal the priorities of the documentarians and archivists involved. Wax disks could only record 6-7 minutes of a song, as the *Federal Cylinder Project* editors note (Gray and Schupman, 1990). Songcatchers might record just one stanza of a song, rather then the whole performance, a fact that highlights that it was the intellectual process of abstracting data from the performance, to be converted to written notation and lyrics, rather than the performance itself, that was most valued for academic folklore purposes. The recordings were data to mine rather than songs to hear. We have also confronted challenges to our efforts in repatriation and exhibition as we began to develop our downloadable "Greatest Hits" of the FCB collection. Because of copyright restrictions on published songs, in the end we may need to limit our final song choices to the purely folkloric – putting us in some cases right back in line with Brown's purity-seeking editors! We have even begun to wonder about the NC Jukebox idea itself as a title and concept, given that the term jukebox didn't come into common parlance until the 1940s, as we learned from *Jukeboxes:An American Social History*, and the music we are sharing was most likely shared at the time either locally or over live radio, as our guest Terry McKinney told us (Segrave, 2002). Our presentism is a both a problem and an acknowledgement of our own historicity and subjectivity.

NC Jukebox is serving as a prototype for future "Archives Alive" projects at Duke University in terms of pedagogical approach, access to primary and secondary source materials from Special Collections, community engagement, and digital platform and exhibition development. As a digital heritage project designed to serve multiple audiences, the digital and onsite exhibition components are being built with an eye towards multiple display formats and locations for a single set of materials within a grown growing database of content. This includes exhibition in a regiona museum organized by McKinney himself (Neufeld, 2015). It also necessitates metadata standards and library infrastructure, a conversation that is ongoing with our Library staff. For the exhibits in Western North Carolina and at the Rubenstein Library we have websites, Omeka exhibits, and interactive web graphics, as well as the downloadable playlist and physical exhibits and listening stations. For the archive, however, we are working with Library and technology partners on a more sustainable and flexible content management system in Drupal that draws content from the more permanent institutional repository solution, which will serve as the substrate for future development as well as, we hope, drive later installations. Our hope is that subsequent generations of students, librarians, and scholars will be able to build upon what we have done in bringing the archives alive.

Bibliography

Brown, F.C. (2016). The Frank C. Brown Collection of North Carolina Folklore; the folklore of North Carolina, collected by Dr. Frank C. Brown during the years 1912 to 1943, in collaboration with the North Carolina Folklore Society : Frank C. Brown Collection of North Carolina Folklore : Free Download and Streaming: Internet Archive. (2016). [online] Internet Archive. Available at: https://archive.org/details/frankcbrowncolleoofran [Accessed 6 Mar. 2016].

Gray, J. and Schupman, E. (1990). *The Federal cylinder project*. Washington: American Folklife Center, Library of Congress.

Graves, Z. (2015). *...and We're Putting it on Wax (The Frank Clyde Brown Collection) - Bitstreams: The Digital Collections Blog*. [online] Bitstreams: The Digital Collections Blog. Available at: http://blogs.library.duke.edu/bitstreams/2015/06/19/and-were-putting-it-on-wax-the-frank-clyde-brown-collection/ [Accessed 6 Mar. 2016].

McKinney, T. (2015). *North Carolina Mountain Music*. Durham, NC. Available at: http://bit.ly/1MAPQRs [Accessed 6 Mar. 2016].

Neufeld, R. (2015). Visiting Our Past: 1930s a Golden Age for music in WNC. [online] *Citizen Times*. Available at: http://www.citizen-times.com/story/life/2015/02/01/visiting-past-golden-age-music-wnc/22705455/ [Accessed 6 Mar. 2016].

Segrave,K. (2002). *Jukeboxes: An American Social History*. London: McFarland.

Child, F., Heiman, M., Heiman, L. and Child, F. (2001). *The English and Scottish popular ballads*. Northfield, Minn.: Loomis House Press.

Duke University. (2016). Archives Alive Initiative | Trinity College of Arts and Sciences. [online] Trinity.duke.edu. Available at: https://trinity.duke.edu/initiatives/archives-alive [Accessed 6 Mar. 2016].

Duke University. (2016). Guide to the Frank Clyde Brown Papers, 1912-1974. [online] David M. Rubenstein Rare Book and Manuscript Library. Available at: http://library.duke.edu/rubenstein/findingaids/brownfrankclyde/ [Accessed 6 Mar. 2016].

Jones, L. and Forbes, J. (1984). *Minstrel of the Appalachians*. Boone, N.C.: Appalachian Consortium Press.

Ritchie, F., Orr, D. and Parton, D. (2014). *Wayfaring strangers: The Musical Voyage from Scotland and Ulster to Appalachia*. Chapel Hill: UNC Press.

U. S. Library of Congress. (1942). Division of music. Archive of American folk song., *Check-list of recorded songs in the English language in the Archive of American folk song to July, 1940. Alphabetical list with geographical index*. Washington, D.C.: Library of Congress, Division of music

The Online Archive “Forced Labor 1939–1945. Memory and History”. A digital Application for Research and Education

Doris Tausendfreund
doris.tausendfreund@cedis.fu-berlin.de
Freie Universität Berlin, Germany

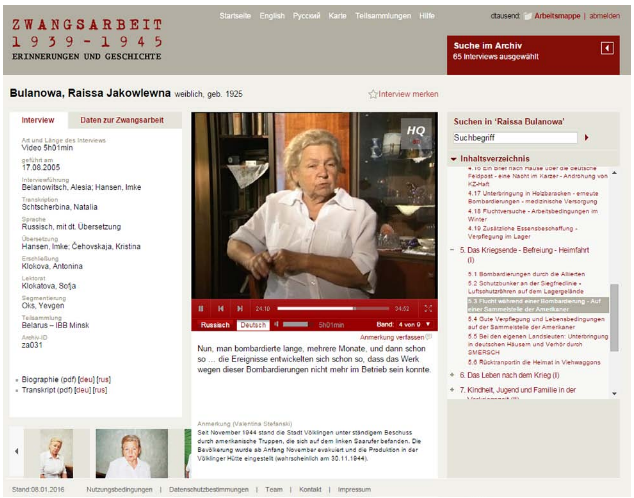
Research

This presentation will focus on the development of the Online-Archive “Forced Labor 1939-1945. Memory and History”.

The collection of narrative interviews was compiled in 2005 and 2006 by the Institute of History and Biography at FernUniversität Hagen. In a joint project, the Foundation “Remembrance, Responsibility and Future”, the Freie Universität Berlin, and the German Historical Museum aim to safeguard and provide easy access to these multilingual audio and video interviews and accompanying materials for research and education.

The online archive contains 583 comprehensive life story interviews with concentration camp survivors, prisoners of war, and “civilian” forced laborers. In 27 countries, mainly in Central and Eastern Europe, 192 video and 391 audio interviews were conducted in the native languages of the witnesses. Each interview is accompanied by additional material: a short biography, a transcript of the interview, a translation of the transcript into German, a table of contents showing the structure of the interview, additional photos and documents, as well as basic biographical information. All content is accessible worldwide for any users who registered with the site.

So far there are no standards on how to document and index Oral History Collections and we will show examples of different approaches which are used at the moment.



Bulanowa, Raissa Jakowlewna weiblich, geb. 1925

Art und Länge des Interviews
Video 5h01min

geführt am
17.08.2005

Interviewführung
Belanowitsch, Alesia; Hansen, Imke

Transkription
Schtscherbina, Natalia

Sprache
Russisch, mit dt. Übersetzung

Übersetzung
Hansen, Imke; Čehovskaja, Kristina

Erschließung
Klokova, Antonina

Lektorat
Klokatova, Sofja

Segmentierung
Oks, Yevgen

Teilsammlung
Belarus – IBB Minsk

Archiv-ID
za031

- Biographie (pdf) [deu] [rus]
- Transkript (pdf) [deu] [rus]

Nun, man bombardierte lange, mehrere Monate, und dann schon so ... die Ereignisse entwickelten sich schon so, dass das Werk wegen dieser Bombardierungen nicht mehr im Betrieb sein konnte.

Anmerkung (Valentina Stefanski)
Seit November 1944 stand die Stadt Völklingen unter ständigem Beschuss durch amerikanische Truppen, die sich auf dem linken Saarufer befanden. Die Bevölkerung wurde ab Anfang November evakuiert und die Produktion in der Völklinger Hütte eingestellt (wahrscheinlich am 30.11.1944).

4.16 Ein Brief nach Hause über die deutsche Feldpost - eine Nacht im Karzer - Androhung von KZ-Haft
4.17 Unterbringung in Holzbaracken - erneute Bombardierungen - medizinische Versorgung
4.18 Fluchtversuche - Arbeitsbedingungen im Winter
4.19 Zusätzliche Essensbeschaffung - Verpflegung im Lager
− 5. Das Kriegsende - Befreiung - Heimfahrt (I)
5.1 Bombardierungen durch die Allierten
5.2 Schutzbunker an der Siegfriedlinie - Luftschutzröhren auf dem Lagergelände
5.3 Flucht während einer Bombardierung - Auf einer Sammelstelle der Amerikaner
5.4 Gute Verpflegung und Lebensbedingungen auf der Sammelstelle der Amerikaner
5.5 Bei den eigenen Landsleuten: Unterbringung in deutschen Häusern und Verhör durch SMERSCH
5.6 Rücktranportin die Heimat in Viehwaggons
+ 6. Das Leben nach dem Krieg (I)
+ 7. Kindheit, Jugend und Familie in der Vorkriegszeit (II)

In this context we will describe our indexing method, with its internal working interfaces and the process involved, as well as the public online application and its functionalities. We will present the different functionalities (content-based indexing, full-text search and an interactive map application) that enable a targeted search that leads directly to individual passages of the interviews.

We will also discuss considerations involved in designing an online platform to avoid the use of the interviews as a mere quotations quarry and instead supports a comprehensive understanding of the whole testimony in its narrative structure and its biographical meaning.

An annotation function will be presented. The function is meant to benefit from the specific knowledge of users to add to the understanding of the interviews.

Finally, the archive has been designed multilingually and runs in German, English, and Russian in order to accommodate the needs of a greater international audience.

This presentation doesn´t focus on a special research problem. Instead it shows a powerful tool which enables academics to work effectively with testimonies to answer their own research questions.

Education

The online-platform aims to support education as well, and there is the option to give an overview of our approaches in this context, too.

We created an online-learning-environment for the use in the classroom which will be available from the beginning of 2016. Based on short biographical films (half an hour) and additional provided material (maps, documents, photographs, additional films) students are asked to work on a number of didactically framed tasks. Most tasks are historical in nature and working with them is useful for the teaching of history. But also for other school subjects exercises are available as for example in language education or religious instructions.

The pupils are asked to write their answers in an online-editor and combine them with selected materials which are easy to import. The results can be saved and printed individually.

- Teachers have additional options and can for example create their own questions based on the films and the materials.
- The software offers versions of the learning environment for different countries. These versions vary in language and content and are drawn up by experts in the respective countries. From the beginning of 2016 a Czech and a German version are available. A Russian version is on the way.
- The online-learning-environment is responsive and can be used with different mobile devices.

Links

http://www.zwangsarbeit-archiv.de
https://lernen-mit-interviews.de/

Bibliography

Andresen, K., Apel, L. and Heinsohn, K. (Eds.) (2015). *Es gilt das gesprochene Wort. Oral History und Zeitgeschichte heute*. Göttingen: Wallstein.

Apostolopoulos, N., Barricelli, M., Possekel, R. and Koch, G. (Eds.) (2016). *Preserving Survivors' Memories. Digital Testimony Collections about Nazi Persecution*. Berlin (forthcoming).

Apostolopoulos, N. and Pagenstecher, C. (Eds.) (2013). *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt*. Berlin: Metropol.

Bothe, A. and Brüning, Ch. I. (Eds.) (2015). *Geschlecht und Erinnerung im digitalen Zeitalter. Neue Perspektiven auf ZeitzeugInnenarchive*. Berlin: LIT Verlag.

De Jong, F., Oard, D. W., Heeren, W. and Ordelman, R. (2008). Access to recorded interviews: A research agenda. *ACM J. Comput. Cultur. Heritage* **1**(1), http://doi.acm.org/10.1145/1367080.1367083

Nägel, V. (2016). Zeugnis – Artefakt – Digitalisat. Zur Bedeutung der Entstehungs- und Aufbereitungsprozesse von Oral History-Interviews. In: Eusterschulte, A., Knopp, S. and Schulze, S. (Eds.), *Videographierte Zeugenschaft. Ein interdisziplinärer Dialog, Weilerswist: Velbrück Wissenschaft* (in print).

Plato, A. von, Leh, A. and Thonfeld, Ch. (2010). *Hitler's Slaves. Life Stories of Forced Labourers in Nazi-Occupied Europe*. New York: Berghahn Books.

Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High Performance Computing, and transforming access to British Library Digital Collections

Melissa Terras
m.terras@ucl.ac.uk
University College London

James Baker
james.baker@sussex.ac.uk
University of Sussex

James Hetherington
j.hetherington@ucl.ac.uk
University College London

David Beavan
d.beavan@ucl.ac.uk
University College London

Anne Welsh
a.welsh@ucl.ac.uk
University College London

Helen O'Neill
helen.oneill@londonlibrary.co.uk
University College London

Will Finley
wafinley1@sheffield.ac.uk
University of Sheffield

Oliver Duke-Williams
o.duke-williams@ucl.ac.uk
University College London

Adam Farquhar
Adam.Farquhar@bl.uk
British Library

How best can humanities researchers access and analyse large-scale digital datasets available from institutions in the cultural and heritage sector? What barriers remain in place for those from the humanities wishing to use high performance computing to provide insights into historical datasets? This paper describes a pilot project that worked in collaboration with non-computationally trained humanities researchers to identify and overcome barriers to complex analysis of large-scale digital collections using institutional university frameworks that routinely support the processing of large-scale data sets for research purposes in the sciences. The project brought together humanities researchers, research software engineers, and information professionals from the British Library Digital Scholarship Department[1], UCL Centre for Digital Humanities (UCLDH)[2], UCL Centre for Advanced Spatial Analysis (UCL CASA)[3], and UCL Research IT Services (UCL RITS)[4] to analyse an open-licensed, large-scale dataset from the British Library. While useful research results were generated, undertaking this project clarified the technical and procedural barriers that exist when humanities researchers attempt to utilize computational research infrastructures in the pursuit of their own research questions.

Overview

The drive in the Gallery, Library, Archive, and Museum (GLAM) sector towards opening up collections data,[5] as well as the growth in data published by publicly-funded research projects, means humanities researchers have a wealth of large-scale digital collections available to them (Lui, 2015, Terras 2015). Many of these datasets are released under open licences that permit uninhibited use by anyone with an internet collection and modest storage capacity. A few humanities researchers have exploited these resources, and their interpretations make claims that change our understanding of cultural phenomena (for example, see Schmidt, 2014; Smith et al., 2015; Cordell et al., 2013; Huber, 2007; Leetaru, 2015). Nevertheless, there remain major barriers to the widespread uptake of these data sets, and related computational approaches, by humanities researchers, which risks diminishing the relevance of the humanities in "big data" analysis (Wynne, 2015). These barriers include:

- fragmentation of communities, resources, and tools;
- lack of interoperability;
- lack of technical skills: "mainstream researchers in the humanities and social sciences often don't know what the new possibilities are" (ibid) and seldom have the technical experience to experiment (Hughes, 2009; Mahony and Pierazzo, 2012).

A common response to this lack of awareness and computational skills is to build web-based interfaces to data[6] or federated services and infrastructures[7]. Whilst these interfaces play a positive role in introducing humanities researchers to large-scale digital collections, they rarely fulfil the complex needs of humanities research which constantly questions received approaches and results, or allow researchers to tailor analysis without being limited by shared assumptions and methods (Wynne, 2013).

Method

We explored the challenges associated with deploying and working with large-scale digital collections suitable for humanities research, using a public domain digital collection provided by the British Library[8]. This 60,000 book dataset covers publication from the 17th, 18th, and 19th centuries, or – seen as data – 224GB of compressed ALTO XML that includes both content (captured using an OCR process) and the location of that content on a page. Using UCL's centrally funded computing facilities[9] we worked from March-July 2015 with RITS and a cohort of four humanities researchers (from doctoral candidates to mid-career scholars) to ask queries that could not be satisfied by search and discovery orientated graphical user interfaces. Working in collaboration we turned their research questions into computational queries, explored ways in which the returned data could be visualised, and captured their thoughts on the process through semi-structured interviews.

Results

We successfully ran queries across the dataset tracking linguistic change, identifying core phrases, plotting and understanding the placing of illustrations, and mapping locations mentioned within core texts. We found that building queries that generate derived datasets from large-scale digital collections (small enough to be worked on locally with familiar tools) is an effective means of empowering non-computationally trained humanities researchers to develop the skill-sets required to undertake complex analysis of humanities data.[10]

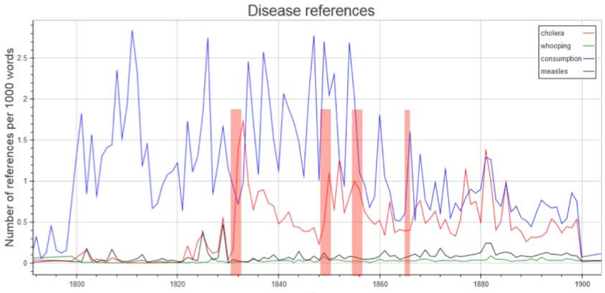


Figure 1: A search for mentions of various infectious diseases (Cholera, Whooping Cough, Consumption, and Measles) across the 60,000 book dataset. We compared the profound spikes for Cholera in the dataset with known data regarding epidemics in the UK (Chadwick, 1842; Wall, 1893) which appear as the bars on the graph, showing a relationship between the first major UK outbreak of Cholera and its appearance within the written record of the time. There are further pronounced spikes 1870s and 1880s: these are not associated with UK epidemics, but there were outbreaks in the US and elsewhere. Identifying the texts that refer to these outbreaks allows us to look more closely at these clusters and to understand the relationship between public health, epidemiology, and the published historical record.

From a technical perspective, this pilot highlighted various sticking points when using infrastructure developed predominantly for scientific research. 224GB is only moderately large by comparison to the scientific datasets UCL RITS usually encounters, but although there are shared assumptions between research infrastructures (adoption of technical standards, and the sharing of tools, approaches and research outputs (Wynne, 2015)) most of the UK's university eScience[11] infrastructure has been constructed specifically to run scientific and engineering simulations, not for search and analysis of heterogeneous datasets. Our task here had a large textual input, a simple calculation, and a small output summary. By comparison, the typical engineering simulation addresses moderately sized numerical input data, runs a long, complicated calculation, and produces a large output. Poor uptake in the arts and humanities (Atkins et al., 2010; Voss, 2010) has meant that these resources have not been optimised for these workloads. The file system and network configuration of Legion – UCL RITS's centrally funded resource for running complex and large computational scientific queries across a large number of cores – did not match the way that the dataset in question was structured (a large number of small zipped XML files).

The complexities associated with redeploying architectures designed to work with scientific data (massive yet very structured) to the processing of humanities data (not massive but more unstructured) should not be understated, and are a major finding of this project. Relevant libraries (such as an efficient XML processor) needed to be installed and optimised for the hardware. Also, the data needed to be transformed to a structure that the parallel file system (Lustre) could address efficiently (that is, fewer, larger files).

Best practice recommendations for comparable projects emerged from this work: the need to build multiple derived datasets (counts of books and words per year, words and pages per book, etc) to normalise results and maintain statistical validity; the necessity of documenting decisions taken when processing data and metadata; and the value of having fixed, definable data for researchers to explain results in relation to (and in turn, the risks associated with iterating datasets). Pointers to how to process the derived datasets were welcomed, but it was at this stage that the researchers were confident to "go it alone" without our support. We also discovered that a core set of four or five queries gave most of the humanities researchers the type of information they required to take a subset of data away to process effectively themselves: for example, keywords in context traced over time; NOT searches for a word or phrase that ignored another word or phrase, etc. As Higher Education Institution (HEI)-based subject librarians regularly handle routine research queries, we contend that training librarians to aid humanities researchers in carrying out defined computational queries via adjustable recipes would improve access to infrastructure, and cut down on the human-resource intensive nature of this approach. In turn, research computing programmers could be invoked as collaborators for their expertise, such as for developing more complex searches beyond the basic recipes.

Conclusion

We successfully mounted large-scale humanities data on high performance computing University infrastructure in an interdisciplinary project that required input from many professionals to aid the humanities scholars in their research tasks. The collaborative approach we undertook in this project is labour intensive and does not scale. Nevertheless, we found many research questions can be expressed with similar computational queries, albeit with parameters adjusted to suit. We recommend, therefore, that HEIs or HEI clusters looking to build capacity for enabling complex analysis of large-scale digital collections by their non-computationally trained humanities research should consider the following activities:

- Invest in research software engineer capacity to deploy and maintain openly licensed large-scale digital collections from across the GLAM sector in order to facilitate research in the arts, humanities and social and historical sciences,
- Invest in training library staff to run these initial queries in collaboration with humanities faculty, to support work with subsets of data that are produced, and to document and manage resulting code and derived data.

Our pilot project demonstrates that there are at present too many technical hurdles for most individuals in the arts and humanities to consider analysing large-scale open data sets. Those hurdles can be removed with initial help in ingest and deployment of the data, and the provision of specific, structured, training and support which will allow humanities researchers to get to a subset of useful data they can comfortably and more simply process themselves, without the need for extensive support.

Bibliography

Atkins, D. E., Borgman, C. L., Bindhoff, N., Ellisman, M., Felman, S., Foster, I. and Heck, A. (2010). RCUK Review of e-Science 2009. Research Councils UK. https://www.epsrc.ac.uk/newsevents/pubs/rcuk-review-of-e-science-2009-building-a-uk-foundation-for-the-transformative-enhancement-of-research-and-innovation/

Huber, M. (2007). The Old Bailey Proceedings, 1674-1834 Evaluating and annotating a corpus of 18th- and 19th-century spoken English. *Studies in Variation, Contacts and Change in English 1: Annotating Variation and Change.* http://www.helsinki.fi/varieng/series/volumes/01/huber/

Hughes, A. (2009). *Higher Education in a Web 2.0 World.* Jisc, http://www.webarchive.org.uk/wayback/archive/20140614042502/http://www.jisc.ac.uk/publications/generalpublications/2009/heweb2.aspx.

Leetaru, K. (2015). *History As Big Data: 500 Years Of Book Images And Mapping Millions Of Books.* Forbes, Tech, http://www.forbes.com/sites/kalevleetaru/2015/09/16/history-as-big-data-500-years-of-book-images-and-mapping-millions-of-books/.

Lui, A. (2015). Data Collections and Datasets, Curated by Alan

Liu. http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244469/Data%20Collections%20and%20Datasets.

Mahony, S. and Pierazzo, E. (2012). Teaching Skills or Teaching Methodology? In Hirsch, B. D. (ed.), *Digital Humanities Pedagogy: Practices, Principles and Politics*, Open Book Publishers, http://www.openbookpublishers.com/product/161/digital-humanities-pedagogy--practices--principles-and-politics.

Schmidt, B. (2014). Shipping maps and how states see. Sapping Attention Blog, http://sappingattention.blogspot.co.uk/2014/03/shipping-maps-and-how-states-see.html.

Smith, D., Cordell, R. and Mullen, A. (2015). Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, **27**(3).

Terras, M. (2015). Opening Access to collections: the making and using of open digitised cultural content. *Online Information Review*, **39**(5): 733–52. http://www.emeraldinsight.com/doi/full/10.1108/OIR-06-2015-0193

Voss, A., Asgari-Targhi, M., Procter, R. and Fergusson, D. (2010). Adoption of e-Infrastructure services: configurations of practice. *Philosophical Transactions of the Royal Society A*. DOI: 10.1098/rsta.2010.0162.

Wynne, M. (2013). The Role of Clarin in Digital Transformations in the Humanities, *International Journal of Humanities and Arts Computing* **7**(1-2): 89-2014.

Wynne, M. (2015). Big Data and Digital Transformations in the Humanities: are we there yet?. *Textual Digital Humanities and Social Sciences Data*, Aberdeen, 21-22 September 2015. http://www.slideshare.net/martinwynne/big-data-and-digital-transformations-in-the-humanities.

Notes

[1] http://britishlibrary.typepad.co.uk/digital-scholarship/

[2] http://www.ucl.ac.uk/dh

[3] http://www.bartlett.ucl.ac.uk/casa

[4] http://www.ucl.ac.uk/isd/services/research-it

[5] http://openglam.org/, an initiative to promote free and open access to digital cultural heritage datasets.

[6] For example, Mining the History of Medicine (http://nactem.ac.uk/hom/) or Language of the State of the Union (http://www.theatlantic.com/politics/archive/2015/01/the-language-of-the-state-of-the-union/384575/).

[7] For example CLARIN (http://clarin.eu/), Common Language Resources and Technology Infrastructure, and DARIAH (https://www.dariah.eu/) Digital Research Infrastructure for the Arts and Humanities.

[8] The British Library has various digital datasets, including (but not limited to) 7m pages of historic newspapers, 1m out of copyright book illustrations, 100,000s of scientific articles, text from over 60,000 books, 1000s of UK theses, and various digitized medieval manuscripts. We chose here just one of its large scale datasets to work with in this pilot phase. For the terms under which the British Library makes collections available, see http://www.bl.uk/aboutus/terms/copyright/.

[9] https://wiki.rc.ucl.ac.uk/wiki/Legion, just one of the High Performance Computing facilities available at UCL for researchers, see http://www.ucl.ac.uk/isd/services/research-it/research-computing.

[10] All code, data, visualisations and other outputs from this pilot project are freely available at https://github.com/UCL-dataspring

[11] For more on the UK's eScience infrastructure, see the work of the eScience Institute, http://www.esi.ac.uk/. Plan-Europe is the Platform of National eScience Centers in Europe (http://plan-europe.eu/). In the United States, the equivalent of eScience is known as Cyberinfrastructure, see the National Science Foundation's guides: http://www.nsf.gov/div/index.jsp?div=ACI.

From Order to Order Switch. Mediating between Complexity and Reproducibility in the Context of Automated Literary Annotation

Bögel Thomas
thomas.boegel@informatik.uni-heidelberg.de
University of Heidelberg, Germany

Evelyn Gius
evelyn.gius@uni-hamburg.de
University of Hamburg, Germany

Janina Jacke
janina.jacke@uni-hamburg.de
University of Hamburg, Germany

Jannik Strötgen
jannik.stroetgen@mpi-inf.mpg.de
Max-Planck Institute for Informatics, Germany

Introduction

In the context of literary studies, which are mainly concerned with the hermeneutic interpretation of literary texts, narratological annotation can be helpful in at least two ways. First, the identification of narrative structures can point to peculiarities of the individual texts that are in need of interpretation, thereby advancing the generation of interpretation hypotheses. Second, since narrative structures can often be detected on the surface level of texts and described intersubjectively, narratological analyses may provide a robust and concrete backing for more comprehensive and complex interpretations.

Against this backdrop, the project heureCLÉA aims at developing a "digital heuristic": a functionality that automatically annotates specific narrative features in literary texts. To achieve this, a corpus of short stories is manually and collaboratively annotated based on a narratological tagset (cf. Gius, 2015; Gius/Jacke, 2015a). The automation is subsequently achieved in a combined approach of rule-based NLP methods and machine learning techniques (cf. Bögel et al., 2015a). However, the automation process is complicated by a specific interdisciplinary conflict: the textual phenomena literary scholars are interested in are often very complex and closely interconnected, which seems to significantly hinder the automation process.

In this paper, we present our way of addressing this problem by way of example: we introduce the basic narratological concept of *temporal order* and its theoretical prerequisites/application conditions; we show how the concept's complexity causes technical issues in the context of automation and how converting *order* to the stripped-down concept of *order switch* significantly enhances the automation results; finally, we explain in which way the new concept is still suited for literary analysis.

Collaborative manual annotation of *temporal* order

The discipline of narratology mainly deals with analyzing the (textual) features typical for narrative representation (cf. Meister, 2014). Narratological text analysis is based on often widely accepted narratological concepts or categories. The project heureCLÉA focuses on the operationalization of a subset of these categories: categories that describe temporal relations between a story and its representation. These categories are: *order* (when does an event happen? – when is it told?), *frequency* (how often does it happen? – how often is it told?), and *duration* (how long does it take to happen? – how long does it take to tell about it?) (cf. Genette, 1980). While these categories are reckoned comparably simple and straightforward in narratology, collaborative manual annotation revealed that they are not. We would like to illustrate this using the example of *order* (cf. fig. 1).

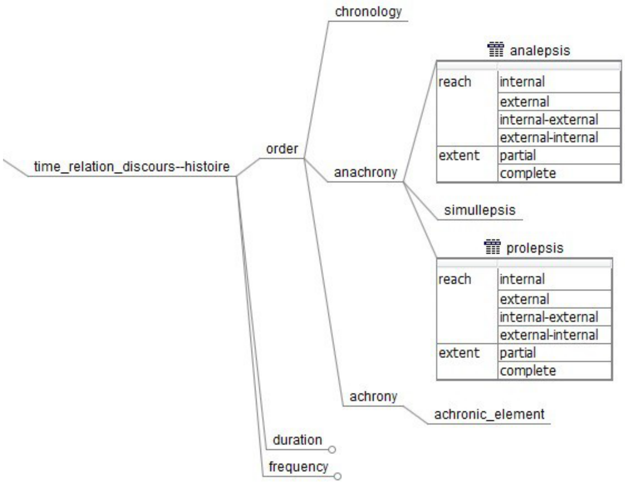
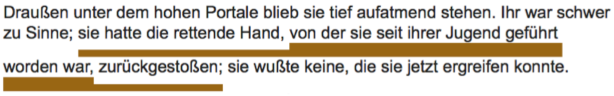


Fig. 1: Tagset order

Basically, the events of a story can either be presented in chronological order or the chronology can be interrupted by "flashbacks" (*analepses* in narratological terms) or "flashforwards" (*prolepses*). Each analepsis and prolepsis can further be qualified according to their reach and extent. Whenever anachronies occur, the whole text passage constituting the anachrony is annotated as either *analepsis* or *prolepsis*. Furthermore, anachronies may be nested: they can contain further anachronies (cf. fig. 2).



Draußen unter dem hohen Portale blieb sie tief aufatmend stehen. Ihr war schwer zu Sinne; sie hatte die rettende Hand, von der sie seit ihrer Jugend geführt worden war, zurückgestoßen; sie wußte keine, die sie jetzt ergreifen konnte.

Fig. 2: Annotation of nested anachronies[1]

The complexity of *order* annotation is significantly increased by the fact that the analysis of *order* showed to be dependent on a different narrative phenomenon: that of narrative levels. Narrative texts can contain embedded narrations, i.e., narrations within narrations. This occurs whenever a character in the story starts telling a story of their own ("new speaker") or when counterfactual passages occur in a narration ("new world") (cf. Ryan, 1991). As should be immediately plausible (at least for ontologically distinct narrative levels), it does not make sense to try and analyze the temporal relation between different narrative levels, i.e., between "actual" and counterfactual events in a story. It thus became necessary to establish an additional round of annotation preceding the annotation of *order*: we first had to identify the embedded narrations in a story, so that temporal order could subsequently be analysed for each narrative level separately.

Automation: from *order* to *order switch*

From a computational linguistic perspective, modeling order phenomena imposes interesting challenges that can be grouped into two types: aspects inherent to the phenomenon and data-specific issues.

Phenomenon-inherent aspects

Regarding characteristics of order phenomena, the aforementioned aspects of nestedness of order poses interesting challenges. As order phenomena are inherently nested, they yield a tree structure of annotations with multiple parent-child relationships. While there are models to formalize and predict tree structures (e.g., in the area of grammars), the prediction is orders of magnitude more complex than the prediction of linear or independent annotations, where complexity in this sense means the amount of training data required to sufficiently model the problem.

In addition, the span of order annotations is highly heterogeneous comprising few tokens as well as multiple paragraphs. Finally, while a sequence classification approach would be suitable to annotate a sequence of tokens representing a specific order, additional aspects of the data at hand impede sequence classification. There is thus no clear annotation target that should be classified by a classifier.

Investigating the data

To assess the annotation quality and thus feasibility of automation, we investigated the primary annotations of order phenomena. Investigating the number of different annotations for the entire training set (21 documents, see below) where two annotators agree with each other – which was the case for 90% of all annotations – revealed that there is an imbalance of annotations with seven times as many analepses (696) than prolepses (98).

This imbalance poses three major problems for statistical modeling and machine learning: *sparsity*, *noise* and class- *imbalance*. Sparsity occurs for annotations that are not well reflected in the data set, such that a classifier cannot find enough evidence to integrate the annotations into its model. This issue is reinforced by noise, meaning inconsistencies in the annotations. One sequence of tokens could either represent a certain order phenomenon or just reflect a change of narrative levels, making it hard for the classifier to learn anything meaningful. Finally, class-imbalance imposes a bias on the classifier, resulting in the phenomenon that the minority class is rarely predicted or even not considered at all.

From order to order switches

Investigating the annotations revealed that sub-sentences serve as boundaries for order switches. Thus, to solve the issues mentioned above, we do not attempt to classify order phenomena directly but instead predict for each sub-sentence whether it introduces a *switch* of the order in the previous (sub-)sentence. While this is, of course, a simplification of the task, it allows us to model the task as a binary classification problem with a clear annotation target and alleviates the issue of sparsity because we do not distinguish between different types of order annotations. To generate training and test data from the original manual annotations of order, we determine all sub-sentences where the order annotation changes, and tag them as order switches.

The resulting annotation statistics are shown in table 1 and indicate that switching from order to order switch increases the number of positive instances in the training set to 1802, meaning that 1802 out of all sub-sentences introduce order changes. Note, however, that the issue of imbalanced data still exists.

Annotation	Mount	percentage
order-switch	1802	12.3%
no switch	12871	87.7%

tab. 1: New training set based on order switches

Evaluation and experiments

Our training set consists of 21 documents from various authors of the 20th century, comprising about 80,000 tokens in total (cf. Bögel et al., 2014). For evaluation, four additional documents were annotated.

Overall, we use 21 features (presented in the appendix) to model order changes. We investigate different aspects of tense (e.g., whether a sub-sentence and the previous (sub-)sentence use the same tense), direct speech, temporal signals (cf. Bögel et al., 2015b), as well as structural features, e.g., paragraph boundaries. Finally, we add features to capture whether the sub-sentence represents a change of narrative levels rather than order.

As mentioned above, the class-imbalance between positive and negative instances remains problematic. To reflect this during the evaluation, we perform randomized re-sampling (cf. Japkowicz/Shaju, 2002) with replacement on the training data which allows us to artificially adjust the spread between two classes. Table 2 contains the evaluation results for different spreads using Random Forests (Breiman, 2001) for classification. The more uniform the distribution of both classes and thus the lower the spread, the better the results. With the best setting (spread = 1:2), we are able to achieve a balanced result with a high F$_1$-score of 81.4%.

Setting	precision	recall	F1
spread=1:6	20.4	14.7	1 7.1
spread=1:3	74.9	7 7.5	76.1
spread=1:2	79.1	83.9	81.4
spread=1:1	76.2	79.5	7 7.8

tab. 2: Evaluation results for different spreads

Overall, the high performance confirms our hypothesis that breaking the complex task of predicting order phenomena into more manageable sub-steps yields promising results.

Nevertheless, we expect that even more complex narrative phenomena can be automatically annotated in the future. As simple narrative concepts have now been tackled successfully, their annotations could be exploited as features to predict more complex phenomena.

Conclusions

Our aim to automate the annotation not only of basic and rather straightforward linguistically encoded temporal aspects like tense and temporal signals (cf. Bögel et al., 2014; Bögel et al., 2015b), but also of more complex phenomena like *order* in heureCLÉA was a long shot. However, by cautiously reducing the concept's complexity in active dialogue between computer scientists and literary scholars, we were not only able to yield good annotation results, but also to end up with a concept that is still of value for literary scholars: while deviations from the chronological presentation of a story cannot be automatically predicted in as much detail as in manual annotation, the automated functionality still serves as a robust heuristic pointing to temporally interesting passages upon which literary scholars can base their in-depth analyses and interpretations. By finding a way to include consideration of narrative levels in the automation, the original *order* concept was not compromised with regards to its conceptual key features. The transformation from *order* to *order switch* is therefore yet another example of successful collaboration between literary scholars and computer scientists in heureCLÉA (cf. Gius/Jacke, 2015b): only a frequent exchange between the involved parties can yield results satisfactory to both sides. We are optimistic that this kind of collaboration has the potential to achieve a functional automated annotation of even more complex narrative phenomena – provided that the phenomena in question are of the kind that their analysis allows for a certain degree of inter-annotator agreement.[2]

Appendix: feature set for order switch prediction

Tense	**tense of target**
tense of target-1	same tense for target&target-1? target contains imperative? target-1 contains imperative
direct speech	target starts direct speech? target-1 starts direct speech? target within direct speech? target-1 within direct speech?
Structural	target occurs after paragraph boundary? target is at beginning/end of sentence? length of target relative to entire sentence
temporal signals	target contains temporal signal? target starts with temporal signal? string of temporal Signac first token of temporal Signac preposition of temporal Signac last token of temporal
Signac narrative levels	target in conjunctive mood? target-1 contains utterance verb

Bibliography

Bögel, Th., Strötgen, J. and Gertz, M. (2014). Computational Narratology: Extracting Tense Clusters from Narrative Texts. *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC'14)*, pp. 950-955, Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/849_Paper.pdf (accessed 02.03.2016).

Bögel, Th. et al. (2015a). Collaborative Text Annotation Meets Machine Learning. heureCLÉA, a Digital Heuristic of Narrative. *DHCommons 1*. http://dhcommons.org/journal/issue-1/collaborative-text-annotation-meets-machine-learning-heurecl%C3%A9-digital-heuristic (accessed 02.03.2016).

Bögel, Th., Strötgen, J. and Gertz, M. (2015b). A Hybrid Approach to Extract Temporal Signals from Narratives. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL'15)*, pp. 106-107, September 20-October 2, Duisburg-Essen, Germany. http://gscl2015.inf.uni-due.de/wp-content/uploads/2015/09/gscl2015-proceedings.pdf (accessed 02.03.2016)

Breiman, L. (2001). Random forests. *Machine learning*, **45** (1): 5-32.

Brunner, A. (2013). Automatic recognition of speech, thought, and writing representation in German narrative texts. *Literary and Linguistic Computing* **28**(4): 563-75.

Genette, G. (1980). *Narrative Discourse*. Oxford: Blackwell.

Gius, E. (2015). *Erzählen über Konflikte: Ein Beitrag zur digitalen Narratologie*. Berlin, München, Boston: De Gruyter.

Gius, E. and Jacke, J. (2015a). Zur Annotation narratologischer Kategorien der Zeit. *Guidelines zur Nutzung des CATMA-Tagsets*. http://www.heureclea.de/guidelines (02.03.2016).

Gius, E. and Jacke, J. (2015b). Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse. *Zeitschrift für digitale Geisteswissenschaften 1*. http://www.zfdg.de/informatik-und-hermeneutik-zum-mehrwert-interdisziplin%C3%A4rer-textanalyse (accessed 02.03.2016).

Japkowicz, N. and Shaju, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis* **6**(5): 429-49.

Meister, J. C. (2014). Narratology. *the living handbook of narratology*. http://www.lhn.uni-hamburg.de/article/narratology (accessed 02.03.2016).

Ryan, M.-L. (1991). *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Bloomington: Indiana University Press.

Storm, T. (1861). *Veronika*. https://textgridrep.de/browse.html?id=textgrid:vtkx.0 (accessed 02.03.2016).

Notes

[1] "Outside under the high gateway she stopped, breathing deeply. Her heart grew heavy; she had [just] pushed back the helping hand by which she had been guided since her youth; she knew none she could grasp now." (Theodor Storm: Veronika).

[2] This precondition may be the critical factor in some automation attempts, e.g., the automated annotation of free indirect discourse that lacks a sufficient amount of reliable indicators (cf. Brunner, 2013). Its determination is rather interpretation-dependent and thus the phenomenon is barely qualified for high inter-annotator agreement.

Printing in a Periphery: a Quantitative Study of Finnish Knowledge Production, 1640–1828

Mikko Tolonen
mikko.tolonen@helsinki.fi
University of Helsinki, Finland

Niko Ilomäki
niko.ilomaki@helsinki.fi
University of Helsinki, Finland

Hege Roivainen
hege.roivainen@helsinki.fi
University of Helsinki, Finland

Leo Lahti
leo.lahti@iki.fi
University of Helsinki, Finland

This paper presents a transparent and quantitative analysis of the overall development of Finnish book production between 1640-1828. The work is based on automated information extraction from library catalogues, and introduces the concept of open analytical ecosystems as a novel research tool for digital humanities. This extends our earlier pilot project on the use of the English Short Title (ESTC) catalogue (https://github.com/rOpenGov/estc). In this new work we focus on Scandinavia, further demonstrating the potential of digitized library catalogues as a valuable resource for digital humanities and reproducible research. We continue our experimental analysis of paper consumption in early modern book production, and provide a practical demonstration on the importance of open-science principles for digital humanities.

Compared to our earlier British analysis (Lahti et al., 2015) we now integrate data across multiple library catalogues from Finland and Sweden. This analysis transcends national boundaries and brings forward key questions in metadata integration such as entry harmonization and duplicate identification. We propose a set of best practices for such tasks in automated large-scale analyses, and exemplify their use in the Scandinavian context. Such pilot project is crucial to later integrate data across the Heritage of the Printed Book database that eventually covers all of early modern Europe. Our emerging data analytical ecosystem supports these goals concretely.

Instead of ready-made standard software, such as Open Refine, Palladio, or similar user-friendly software, we have developed a set of custom tools in the R statistical programming environment to combine automation with full flexibility and access to state-of-the-art data analysis and visualization algorithms. An important contribution in comparison with related earlier work, for example Kalev's GDELT (http://blog.gdeltproject.org/mapping-212-years-of-history-through-books/), is that we have drastically refined the metadata, for instance by harmonizing synonymous entries and by enriching the data with external information such as name-gender mappings and geographical information. The bibliographic metadata in national library catalogues follow international standards and, as we demonstrate, the fully open source computational data analysis tools introduced within this project are immediately relevant and widely applicable in further studies based on library catalogue metadata.

We focus on the extraction and statistical analysis of library catalogue metadata to study the emergence and development of public discourse in Finland (1640–1828). The main data source for our analysis is Fennica, Finnish National Bibliography (https://github.com/rOpenGov/fennica). This is complemented with further metadata and content analysis of Finnish newspapers and journals and material from Sweden, from the Kungliga collection, Stockholm (https://github.com/rOpenGov/kungliga), and include comparisons with further library catalogue material from other countries as well. The analysis allows us to provide concrete, quantitative figures on publication activity, places, and topics and compare these to political, technological, and social ruptures. The quantitative analysis of print culture will allow us to study how the development of Finnish book, newspaper and journal production compares to European trends.

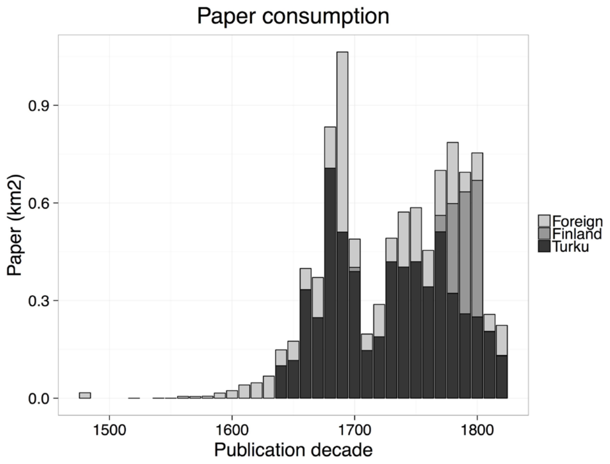


Fig. 1. Paper consumption of documents recorded in Fennica until 1828 by place of publication (Turku, other places in Finland and elsewhere including Sweden)

It is not enough, however, to see European public discourse by combining nationally organized knowledge. The hypothesis is that the European map of knowledge production will have local flavors in different corners of Europe. The aim should thus be to integrate data across library catalogues to analyze different streams of influence and varying regional perspectives and uncover potential asymmetries that may have guided intellectual life. The comparison between neighboring countries also allows for the detection of local publishing networks in the Baltic Sea region.

Our aim in this paper is particularly to study the development of publishing houses in Finland and their spread from Turku to other parts of Finland. We will also identify overlooked moments of transformation in public discourse in Finland by blending historical and computational approaches. The research undertaken will reflect on how social change and public discourse are intertwined, and how cultural, institutional, legal and technological changes are reflected both in publication metadata and the textual content of the publications. In terms of the historical timeframe, our study begins with the founding of the first Finnish press at the Academy of Turku in 1640, tracks the overall publishing history of the country until 1828 when Helsinki starts to play a major role in Finland.

Public discourse in Finland has been largely approached from the perspective of the breakthrough of the Finnish language, the role of elite discourse at the university, early Swedish-language newspapers, and book history. We combine these perspectives, and further analyze how language-barriers, elite culture and popular debate, as well as different publication channels interacted. Large-scale quantitative analysis of library catalogues opens novel opportunities to characterize the general impact of the turn from Swedish to Russian rule in early nineteenth-century regarding public discourse in Finland. Previous historical research on the development of civility in nineteenth-century Finland has lacked appropriate quantitative tools to take an objective 'bird-eye' view of these complicated and crucial transformations. Questions of how, for instance, the establishment of the university in Turku/Åbo (1640), the introduction of freedom of print (1766), the formation of a Finnish Grand Duchy in the Russian Empire (1809–1812), the changes in the enforcement of censorship, the decision to transform Helsinki into a capital city (1819), the lack of estate representation in the Grand Duchy, and the slow emergence of a Finnish written language resonated in publication practices are explored from a quantitative perspective.

Our open data analytical ecosystems provide powerful and flexible data analytical tools that can best serve the needs of genuinely data-intensive research, in contrast to traditional point-and-click interfaces that are suitable for simple query tasks but not designed for fully transparent, reproducible and automated large-scale algorithmic data mining. The open source ecosystems will also enable new collaboration models around digital data collections that are now becoming increasingly available for research and other purposes. This emphasis on transparent and collaborative methodology, already wide-spread in other fields of computational science, sets the context for our work within digital humanities. Others can benefit from the new tools and the libraries from the refined data sets. The data analytical algorithms, including data extraction, statistical analysis, summarization and reporting, will be are released in full detail within a unified open source ecosystem in Github (http://github.com/rOpenGov/fennica), where all steps from raw data to the final results can be traced back and improved further. In this sense, our emphasis on open data analytical process and collaboration model is different from Anderson's and Blanke's discussion on digital humanities ecosystems (Anderson and Blanke, 2012), which focuses on the role of the community of researchers.

This paper continues an ongoing trend of quantitative analysis of publishing history (Moretti, 2013; Towsey et al., 2015). While reuse of library catalogue data has been discussed in recent digital humanities scholarship (for example, Prescott, 2013 and Bode and Osborne, 2014), large-scale library catalogues represent a so far underestimated research resource, containing systematic information on publication activity over years, language barriers, genres, geographical regions and other variables in which the evolution of the public sphere is reflected. Moreover, our work complements the distant reading and related concepts discussed by Bode and Moretti by introducing concrete algorithms and proposing a collaborative development model. Lincoln Mullen has used a related approach to analyze historical texts (http://lincolnmullen.com/#software).

Our ultimate aim is to develop algorithms to extract, harmonize and integrate relevant metadata across the different European library catalogues, regardless of language. This research data will be further enriched by complementing it with data from auxiliary sources, such as linked data on person records, biographies of partner organizations, ontologies and other assets. The enriched information is then used to formulate statistical summaries and systematic quantitative comparisons, as well as interactive and dynamic visual representations of the publication activity, topics and geographical variation and of their evolution over time. To demonstrate the efficiency of this approach, we quantify the importance of Turku in the Finnish publication landscape during 1640-1828.

Bibliography

Anderson, S. and Blanke, T. (2012). Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems. *Historical Social Research*, **37**: 147-64.

Bode, K. and Osborne, R. (2014). Book History from the Archival Record. In Leslie Howsam (Ed.), *The Cambridge Companion to the History of the Book*. Cambridge University Press, pp. 219-36.

Lahti, L., Ilomäki, N., Tolonen, M. (2015). A Quantitative Study of History in the English Short-Title Catalogue (ESTC) 1470-1800. *LIBER Quarterly*, **25**(2): 87-116.

Moretti, F. (2013). *Distant Reading*. Verso books.

Prescott, A. (2013). Bibliographic records as humanities big data.

Big Data IEEE International Conference 2013: Conference Abstracts. Silicon Valley, pp. 55-58.

Towsey, M., Bode, K. Burrows, S. et al. (2015). Remapping Cultural History: Digital Humanities, Historical Bibliometrics, and the Reception of Print Culture. *Digital Humanities 2015 Conference*, University of Western Sydney.

Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930

Peer Trilcke
trilcke@phil.uni-goettingen.de
University of Göttingen, Germany

Frank Fischer
frank.fischer@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Mathias Göbel
goebel@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Dario Kampkaspar
kampkaspar@hab.de
Herzog August Library Wolfenbüttel, Germany

Approach

Decades ago, alongside more traditional structuralist paradigms that were largely based on linguistic theorems (Lotman 1972, Titzmann 1977), literary studies began to undertake structural analyses based on empirical sociology, in particular the social network analysis. Structure was no longer solely defined by semantic relations (such as opposition or equivalence), but by social interactions, too (Marcus 1973; Stiller, Nettle and Dunbar 2003; de Nooy 2005; Stiller and Hudson 2005; Elson, Dames and McKeown 2010; Agarwal et al., 2012). In the context of the Digital Humanities, this kind of approaches has gained a new dynamic in shape of a dedicated literary network analysis (Moretti 2011; Rydberg-Cox 2011; Park, Kim and Cho 2013; Trilcke 2013). This method is based on the analysis of bigger literary corpora (i.e., quantitative data) and promises insights into the history of literature as well as generic characteristics of literary texts. In our project, "dlina. Digital LIterary Network Analysis", we already developed a workflow for the extraction, analysis and visualisation of network data from dramatic texts built on basic TEI markup (Fischer, Kampkaspar, Göbel, Trilcke, 2015). This paper will present results of our analysis of the network data gathered so far and discuss them in the light of current theories in the field of social network analysis.

Data Collection and Analysis

Our current corpus comprises 465 German-language dramas (from 1730 to 1930), the better part of the Digitale Bibliothek corpus contained in the TextGrid repository (https://textgridrep.de/). The structural data crucial for the network analysis of these dramas (segmentation, character identification, etc.) was revised manually in a rule-based process to straighten out issues with the OCR and TEI tagging. We also had to level out philological peculiarities that would otherwise distort our results (such as different names for identical figures or groups of characters like 'Both' or 'All'). All the structural data is stored in an XML format we especially developed for that purpose (DLINA format). Network visualisation and network-value calculation has been automated (via Python and, alternatively, JavaScript to facilitate a direct embedding of our results into webpages). The scripts are fed with the data stored in DLINA files. In addition to graphs and simple network values that globally describe networks (like network size, density, average degree, average path length, clustering coefficent), we also calculate centrality values for the characters of each play (like degree, average distance, closeness centrality, betweenness centrality). In addition, we most recently implemented the calculation of random graphs based on the observed drama networks. All data and visualisations are freely available online on the project website (https://github.com/dlina and https://dlina.github.io/linas/).

Evaluation, Part I: History of Drama

The diachronic extension of our corpus over 200 years of German literary history allows the observation of larger developments in the structural composition of dramatic texts (we outlined some reflections on this in a blog post: https://dlina.github.io/200-Years-of-Literary-Network-Data/). Values referring to networks as a whole will be broached (incl. network size, density, average degree; as an example, we put average-path-length values by decades in Fig. 1), as will be character-related values for each character of each play (centrality measures, primarily) providing information on the distribution of the personae dramatis or their division into 'central' and 'less central' characters. These values will lay the groundwork for the discussion of some global hypotheses of literary history. We will discuss, firstly, the extent to which we can observe a differentiation of the structural composition of drama at the end of the 18th century on the basis of network analysis values: Such a differentiation is to be expected given the coexistence of 'closed' drama (following the doctrines of French classicism) and 'open' drama (mostly influenced by Shakespeare). Secondly, we will discuss some common literary periodising hypotheses (originating from structuralism, social history, or other directions). We will have a closer look at correlations between our network data and well-established traditional periodisations.

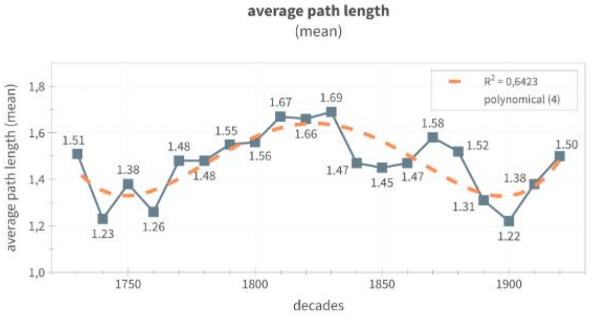


Fig. 1: Average path length by decades (mean)

Evaluation, Part II: Types of Drama

The data raised so far shows how very differently theatre plays were structured in the focal period. Traditional literary studies have developed various typologies for such different types, the most popular in German studies being Volker Klotz's subdivision into 'open' and 'closed' drama. We want to build on this kind of typological impulse and propose a method as to how certain types of structural composition can be distinguished by means of network analysis (and also placed in their historic context). With this proposal we want to take up reflections from research on so-called small-world networks. This branch of research assumes that the values of empirically collected networks often differ significantly from those of corresponding random networks (e.g., graphs generated with the Erdős–Rényi model). Following the approach of Stiller, Nettle and Dunbar 2003, but relying on a much larger set of texts, we investigate the plays in our corpus with regard to their small-world properties (clustering coefficient, average path length, node degree distribution). The results show that there are just a few plays that meet all the criteria (a total of five plays, i.e., just about one percent of the corpus) – see figs. 2.1 to 2.5.

These findings will give us a deeper understanding of different types of structural composition. We shall first direct our attention to forms of networks that – unlike dramas with small-world properties – occur much more frequently in our corpus. Eventually, we will discuss structural characteristics of drama networks exhibiting properties exactly opposite to the properties of small-world dramas (e.g., reverse power-law form in the node degree distribution). It will also be discussed in this context whether we can contrast the strong hierarchical type of small-world drama with an anti-hierarchical type.

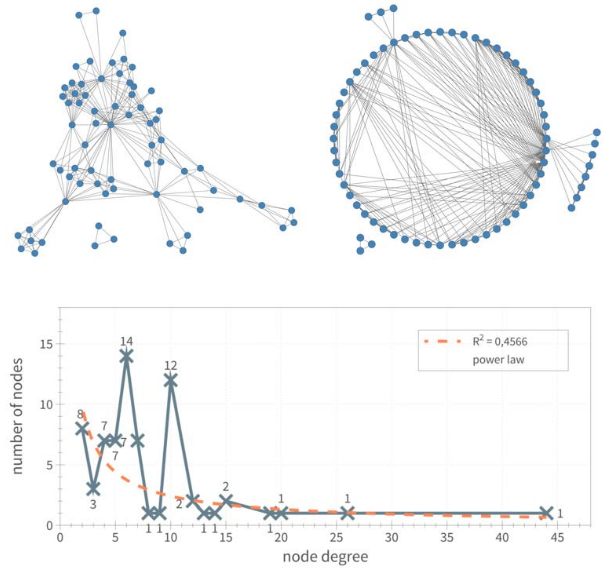


Fig. 2.1: Goethe, “Götz von Berlichingen” (1773): Spring Embedder Layout, Circular Layout, Node Degree Distribution

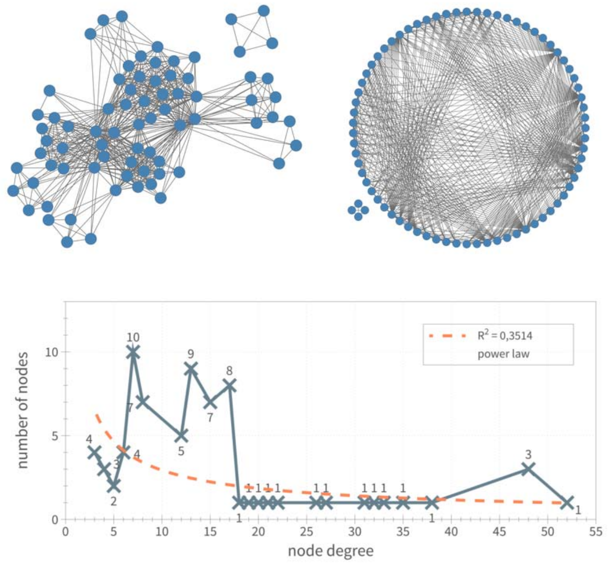


Fig. 2.2: Arnim, “Jerusalem” (1811): Spring Embedder Layout, Circular Layout, Node Degree Distribution

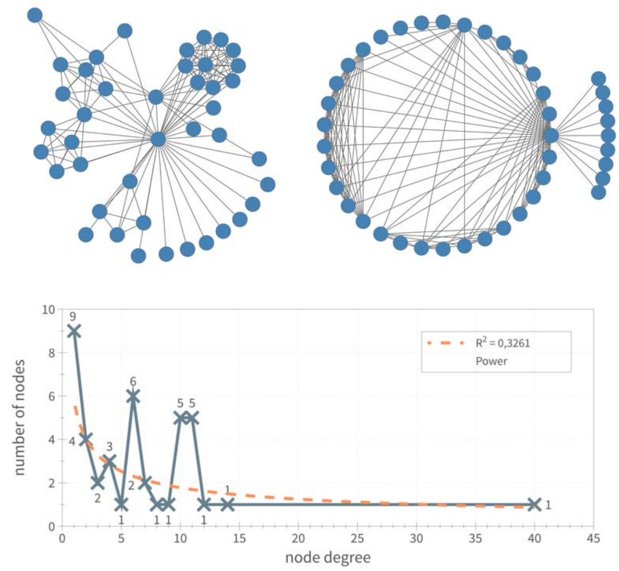


Fig. 2.3: Soden, “Doktor Faust” (1797): Spring Embedder Layout, Circular Layout, Node Degree Distribution

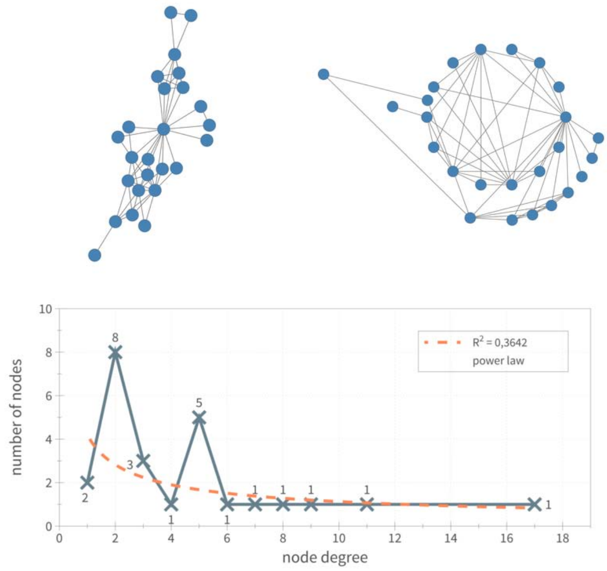


Fig. 2.5: Raimund, “Der Barometermacher” (1823): Spring Embedder Layout, Circular Layout, Node Degree Distribution

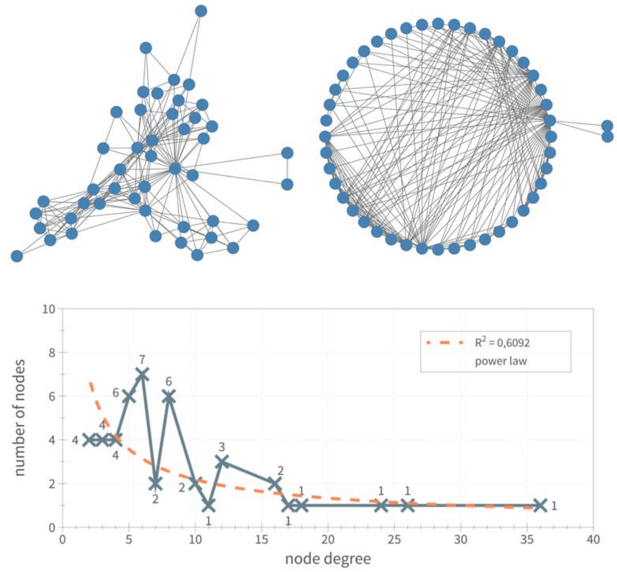


Fig. 2.4: Nestroy, “Der böse Geist” (1833): Spring Embedder Layout, Circular Layout, Node Degree Distribution

Bibliography

Réka, A. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**: 47–97.

Agarwal, A. et al. (2012). Social Network Analysis of Alice in Wonderland. *Proceedings of the Workshop on Computational Linguistics for Literature*, Montréal: 88–96.

de Nooy, Wouter (2006). Stories, Scripts, Roles, and Networks. *Structure and Dynamics* **1**(2) http://escholarship.org/uc/item/8508h946#page-1 (accessed 4 March 2016)

Elson, D. K., Dames, N.and McKeown, K. R. (2010). Extracting Social Networks from Literary Fiction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, pp. 138–47.

Fischer, F., Kampkaspar, D.; Göbel, M. and Trilcke, P. (2015). Digital Network Analysis of Dramatic Texts. *DH2015*, script: https://dlina.github.io/Our-Talk-at-DH2015/, slides: https://dlina.github.io/presentations/2015-sydney/sydney.html (accessed 4 March 2016).

Klotz, Volker (1960). *Geschlossene und offene Form im Drama*. München.

Lotman, Jurij M. (1972). *Die Struktur literarischer Texte*. München.

Marcus, Solomon (1973). *Mathematische Poetik*. Frankfurt/M.

Moretti, Franco: Network Theory, Plot Analysis. *Stanford Literary Lab Pamphlets*, **No. 2**, 1 May 2011, http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf (accessed 4 March 2016).

Park, G.-M., Sung-Hwan, K. and Cho, H.-G. (2013). Structural Analysis on Social Network Constructed from Characters in Literature Texts. *Journal of Computers* **8**(9): 2442–47, http://ojs.academypublisher.com/index.php/jcp/article/view/jcp080924422447/7672 (accessed 4 March 2016).

Rydberg-Cox, J. (2011). Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science,* **1**(3), https://letterpress.uchicago.edu/index.php/jdhcs/article/view/86/91

Stiller, J., Nettle, D. and Dunbar, Robin I. M. (2003). The Small World of Shakespeare's Plays. *Human Nature* **14**: 397–408.

Stiller, J. and Hudson, M. (2005). Weak Links and Scene Cliques Within the Small World of Shakespeare. *Journal of Cultural and Evolutionary Psychology* **3**: 57–73.

Titzmann, M. (1977). *Strukturale Textanalyse. Theorie und Praxis der Interpretation.* München.

Trilcke, P. (2013). Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. Ajouri, Philip; Mellmann, Katja and Rauen, Christoph (eds.): *Empirie in der Literaturwissenschaft*. Münster, pp. 201–47.

A literary rat race

Karina van Dalen-Oskam
karina.van.dalen@huygens.knaw.nl
Huygens ING, The Netherlands / Universiteit van Amsterdam,
The Netherlands

Introduction

'A hype already', was the Friday 28 January 2011 headline on the first page of the Dutch newspaper *NRCnext*. 'It's called the literary rat race K2, the simultaneous publication of Herman Koch's *Zomerhuis met zwembad* and Kluun's *Haantjes*'. The front-page story continues with literary critic Arjen Fortuin presenting an analysis of the two novels that were published the week before, both written by well-known Dutch authors and with the amazing first print runs of 80,000 (Kluun) and 100,000 (Koch) copies. These two novels are from totally different authors, who began their careers on opposite sides of the literary spectrum. However, Fortuin states, they seem to be converging. Koch started out as a 'literary' author not selling very well, but with his last book before *Zomerhuis met zwembad* (*Summerhouse with Swimming Pool*), *Het diner* (2009, translated into English as *The Dinner*, 2012) he turned to a wider audience, thus - according to literary critics - severely damaging his literary reputation. Kluun (the one-word alias of Raymond van de Klundert) started out as writing popular fiction with no literary pretentions at all with his much read but openly despised *Komt een vrouw bij de dokter* (2003, translated as *Love Life* (2007), under the name Ray Kluun). Quite unexpectedly, on 23 January 2011, Kluun's new novel *Haantjes* (which could be translated as 'Alpha-males') got a positive review from prominent literary critic Arjan Peters in *de Volkskrant*. Fortuin found this an additional reason to compare the two novels. His conclusion is: 'It's an uneven literary match – Kluun plays in a lower league – but the commercial battle of K2 could be a close tie – although here Koch also seems to have the best chances: the Alpha-males are in fact very light-weight'.

The Riddle of Literary Quality

Both novels are on the list of 401 novels analyzed in the project *The Riddle of Literary Quality* (http://literaryquality.huygens.knaw.nl/). The aim of the project (running until 2017) is a stylistic analysis of novels in Dutch and to compare this analysis with readers' opinions. The corpus was based on a list of most sold and most lent titles in The Netherlands from 2010 to 2012, excluding titles first published in Dutch before 2007. It includes Dutch originals (such as the novels by Kluun and Koch) and translations. The list contains a lot of genre fiction such as thrillers, 'literary thrillers', and chick-lit, and many titles the publishers categorized as 'literary novels', among which the K2 titles (Koch's *Zomerhuis met zwembad* and Kluun's *Haantjes*). In 2013, in an online survey titled *Het Nationale Lezersonderzoek* ('The National Reader Survey') we asked a wide audience of readers to indicate which of the 401 novels they had read, and for a smaller set of these novels how they rated them on two scales: one on the scale of literariness, from 1 (not literary at all) to 7 (highly literary) and one on the scale of general quality, from 1 (very bad) to 7 (very good). In total, 13,782 respondents completed the survey. By combining these data with an analysis of stylistic characteristics of the novels we expect we can discover which textual features may play a role in the current Dutch conventions of literariness. In this paper I will compare the K2 novels both in an exploratory analysis of the survey results on the scale of literariness and in stylometric analysis. Do the opinions of the K2 readers agree with the stylometric picture we get?

Zooming in on style

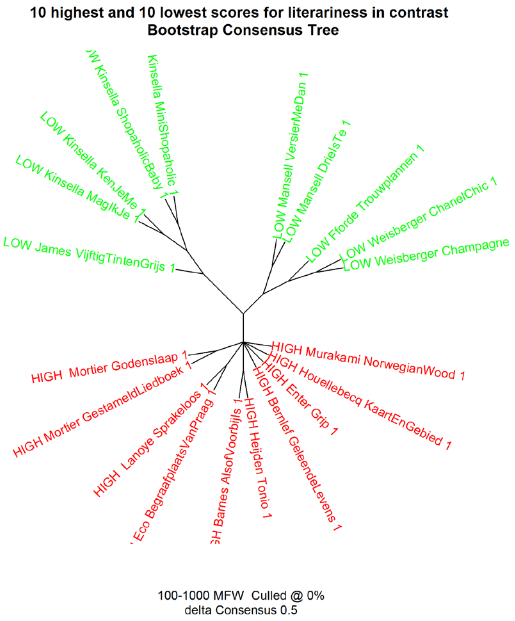

10 highest and 10 lowest scores for literariness in contrast
Bootstrap Consensus Tree
W Kinsella ShopaholicBaby 1
Kinsella MiniShopaholic 1
LOW Kinsella KenJeMe 1
LOW Kinsella MagIkJe 1
LOW James VijftigTintenGrijs 1
LOW Mansell VersierMeDan 1
LOW Mansell DrielsTe 1
LOW Fforde Trouwplannen 1
LOW Weisberger ChanelChic 1
LOW Weisberger Champagne
HIGH Mortier Godenslaap 1
HIGH Mortier GestameldLiedboek 1
HIGH Lanoye Sprakeloos 1
Eco BegraafplaatsVanPraag 1
H Barnes AlsofVoorbijls 1
HIGH Heijden Tonio 1
HIGH Bernlef GeleendeLevens 1
HIGH Enter Grip 1
HIGH Houellebecq KaartEnGebied 1
HIGH Murakami NorwegianWood 1
100-1000 MFW Culled @ 0%
delta Consensus 0.5

Fig. 1

'Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively' (Herrmann et al. 2015). For the K2 case, I use the Stylo package in R (Eder et al. 2013). This R package is mainly used for authorship attribution. It compares texts based on the frequencies of a range of words or characters. Since words give more insight into the texts themselves than characters, Stylo can also be used for literary analysis beyond verifying authorship. In Fig. 1 I used Stylo to measure the distance between the ten 'literary' novels by male authors with the highest scores for literariness and the ten novels that got the lowest scores. The bootstrap consensus tree (a harmonization of cluster analyses) was based on the most frequent 100, 200, 300 etc. words (MFW) until 1000 MFW. There is a complete distinction between the "HIGH" and the "LOW" group.

However, the ten novels with the highest scores are all written by men and the ten with the lowest scores by women. The "HIGH" novels are labelled by the publishers as 'literary novel' and the "LOW" ones are mostly marketed as genre fiction such as chick-lit. The graph therefore probably does not distinguish literary quality but genre.

I now zoom in on the 'literary novels'. The Riddle corpus contains 96 titles labeled as 'literary novel' written by a male author or by only male co-authors, and 66 by a female author or only female co-authors. For the whole set of 401 novels, 191 are written by male and 196 by female authors or co-authors, which shows female authors are not underrepresented in the corpus. In Fig. 2 all 'literary novels' are categorized according to the mean scores for level of literariness. The general trend seems to be that the respondents see the female authors 'playing in a lower league' than the male authors. Riddle-PhD-students Corina Koolen and Kim Jautze will deal with gender issues in detail in their dissertations.

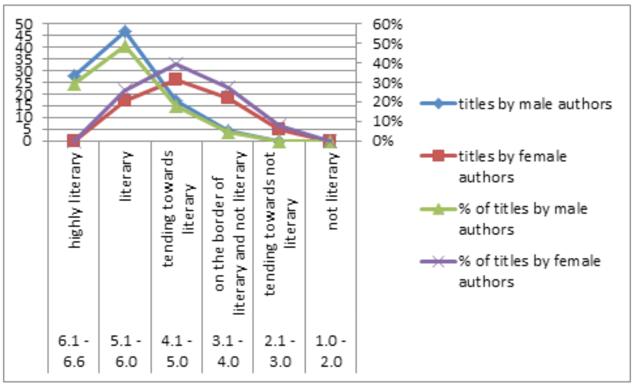


Fig. 2 Literary novels (originally Dutch and translated into Dutch)

I will leave the gender topic to my PhD-students to publish about, and I will for now limit my presentation to an analysis to novels written by male (co-)authors. The lowest score on the level of literariness was 3.1, for two titles written by the Dutch author (and sports reporter) Mart Smeets. Kluun's novel *Haantjes* is directly above these two novels, with a score of 3.5. *Zomerhuis met zwembad* clearly did better. It got a mean score of 5.1. The highest score is 6.6, for Julian Barnes' *Alsof het voorbij is* (*The sense of an ending*). Where Barnes' novel ranks first on the list of 96 novels, the new Koch ends up at rank 74 and the new Kluun at 94. The respondents of *The National Reader Survey* thus agree with Fortuin that Koch ranks higher on the scale of literariness. From the list of 96 novels I select thirty novels: ten with the highest (including Barnes), ten with the lowest (including Kluun), and ten with intermediate scores for literariness (including Koch). If we use the same settings in Stylo as above, the 30 novels end up as shown in Fig. 3.

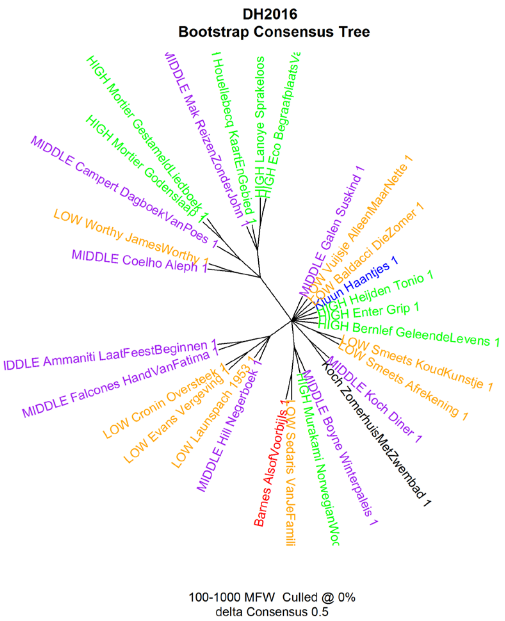


Fig. 3

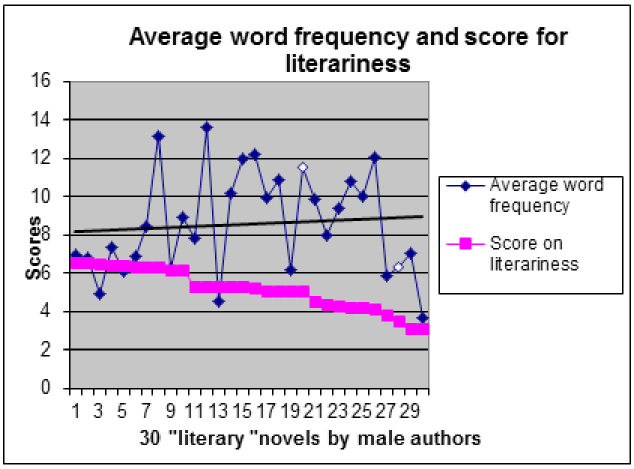


Fig. 4

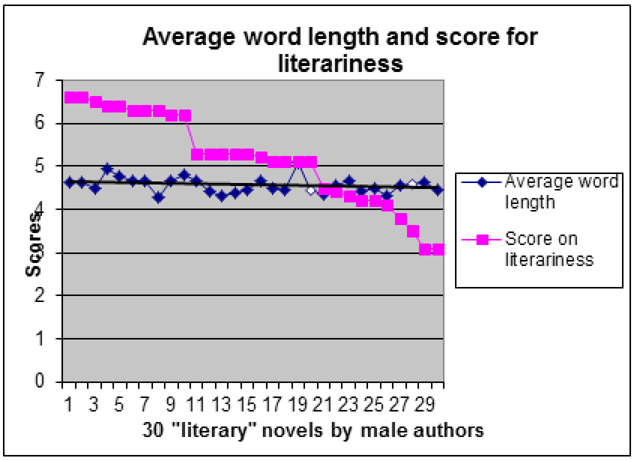


Fig. 5

The results are not very clear. The groups "HIGH", "MIDDLE", and "LOW" do not have their own specific clusters. For now, we can conclude that literary quality for this corpus does not reside in shared word frequency patterns. We are currently gathering as many measures of linguistic features as possible to apply these to the selected corpus. One of the assumptions to test is whether the scores for literariness correlate with features that relate to linguistic complexity. A suite of tools for Dutch is currently being developed. A simple test using HyperPo does show that some features normally related to the level of difficulty of a text need further inspection (Fig. 4, 5, 6) (http://tapor1.mcmaster.ca/~sgs/HyperPo/).

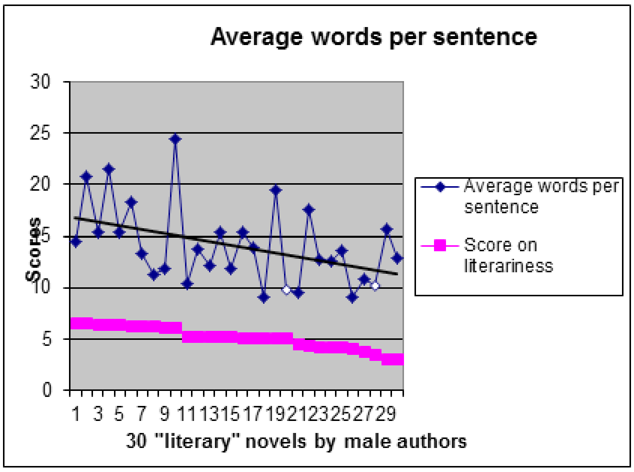


Fig. 6

In each of these figures, the 30 novels are represented on the x axis arranged from highest score for literariness to lowest. Koch is at data point 20 and Kluun at 28. Fig. 4 shows that the scores for literariness display a trend that is opposite to the average frequency of words – this suggests that for this corpus, lexical density scores perhaps play a role in what makes a novel literary or not. However, Kluun's score below the trend line is close to that of the top-5, while Koch is on the opposite side of the trend line. Fig. 5 highlights that mean word length does not resonate with scores for literary quality. Fig. 6, however, shows a clear (statistically significant) trend of literary score and average words per sentence. Here, Koch and Kluun do not differ very much. This sneak preview (which will be tested with tools finetuned for Dutch) shows a further analysis of complexity issues is promising. Riddle PhD student Andreas van Cranenburgh is working on syntactic markers of literariness, and all three PhD students have looked into topic variation in the Riddle corpus (see the abstract they have submitted for DH2016).

Conclusion

So how do the novels by Kluun and Koch compare? The match is still undecided. We need an analysis of many more linguistic features before we can draw conclusions. This analysis did show in which directions to look. For now, stylistically they are playing in the same league.

Bibliography

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: Digital Humanities 2013: Conference Abstracts. University of Nebraska--Lincoln, NE, pp. 487-89.

Fortuin, A. (2011) 'Mogen wij nog haantjes zijn? Tweestrijd van de moderne man is het gemeenschappelijke thema van Kluun en Koch'. [Can we still be alpha-males? Dilemma of the modern man is the shared theme of Kluun and Koch.] In: *ncrnext*, Friday 28 January 2011, p. 4-5. http://www.nrc.nl/next/2011/01/28/mogen-wij-nog-haantjes-zijn-11991928

Herrmann et al. (2015). J. Berenike Herrmann, Karina van Dalen-Oskam, Christof Schöch, Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory* 2015; 9(1): 25–52.

Kluun. (2003) *Komt een vrouw bij de dokter*. Uitgeverij Podium.

Kluun, R. (2007) *Love life*. Pan UK.

Kluun. (2011) *Haantjes*. Amsterdam: Uitgeverij Podium.

Koch, H. (2009) *Het diner*. Amsterdam: Anthos.

Koch, H. (2011) *Zomerhuis met zwembad*. Amsterdam: Ambo|Anthos.

Koch, H. (2012) *The dinner*. Faber & Faber.

Koch, H. (2014) *Summerhouse with swimming pool*. Atlantic Books.

Peters, A. (2011) 'De nieuwe Kluun mag er wezen. Hoogstaand amusement in Haantjes van Kluun. Of: hoe Stijn van Diepen goud geld dacht te verdienen aan de Gay Games in Amsterdam' [The new Kluun is splendid. High quality entertainment in *Haantjes* by Kluun. Or: how Stijn van Diepen expected to make big money at the Gay Games in Amsterdam.] In: *de Volkskrant* 23 January 2011, http://www.volkskrant.nl/boeken/de-nieuwe-kluun-mag-er-wezen~a1827236/

Taalportaal: A New Tool For Linguistic Research

Ton van der Wouden
ton.van.der.wouden@meertens.knaw.nl
Meertens Instituut, Netherlands, The

1 Introducing the Taalportaal

The Taalportaal project aims at the development of a comprehensive and authoritative digital scientific grammar for Dutch and Frisian, the two official (oral) languages of the Netherlands, in the form of a virtual language institute. The Taalportaal is built around an interactive knowledge base of the current grammatical knowledge of Dutch and Frisian. The Taalportaal's prime intended audience is the international scientific community, which is why the language used to describe the language facts is English.

The Taalportaal provides an (almost) exhaustive collection of the currently known data relevant for grammatical research, as well as an overview of the established insights about these data. This is an important step forward compared to presenting the same material in the traditional form of (paper) handbooks. For example, the three subdisciplines syntax, morphology and phonology are often studied in isolation, but by presenting the results of these sub-disciplines on a single digital platform and internally linking these results, the Language Portal contributes to the integration of the results reached within these disciplines.

Technically, the Taalportaal is an XML-database, organized as DITA-topics (cf. https://en.wikipedia.org/wiki/Darwin_Information_Typing_Architecture), that is accessible via the Internet using any standard internet browser. Organization and structure of the linguistic information are reminiscent of, and to a certain extend inspired by, Wikipedia and comparable online information

Sources (but without the anarchy of Wikipedia). The project is a collaboration of the Meertens Institute, the Fryske Akademy, the Institute of Dutch Lexicology and Leiden University, funded, to a large extent, by the Netherlands Organisation for Scientific Research (NWO) (Landsbergen et al. 2014).

Besides the grammar modules, the portal contains an ontology of linguistic terms (recast recently in the CLARIN Concept Registry (Schuurman (2015), cf. https://www.clarin.eu/ccr) and an extensive bibliography. As of January 2016, the first release of the Taalportaal is online via http://www.taalportaal.org.

2 Enriching the Taalportaal with links to linguistic resources

The Taalportaal database has been enriched with links to on-line linguistic resources. Links between a descriptive grammar and a linguistically annotated corpus are valuable for various reasons. Illustrating a given construction with corpus examples may help to get a better understanding of the variation of the construction and the frequency of these variants, as well as give insight into the lexical items that occur most often in the pertinent construction. Corpus data may also convince a reader that a given variant really occurs in (well-formed) text. Finally, corpus data may also yield occurrences of constructions judged ungrammatical by the authors of the descriptive grammar for reasons such as prescriptivism or theoretical bias. The possibility of enriching a grammar with links to on-line linguistic resources is thus a unique selling point of digital grammars vis-a-vis old school paper grammars (van der Wouden et al. 2015).

Luckily, there is no lack of linguistic resources for Dutch that are useful for this purpose (Frisian is a slightly different matter), e.g. the (syntactically annotated part of the) Corpus of Spoken Dutch (manually verified syntactic annotation for 1M words of speech) (van der Wouden et al. 2002b; Schuurman et al. 2003), the Lassy Small treebank (manually verified syntactic annotation for 1M words of text from various genres) and the Lassy Large treebank (700M words of text, automatic syntactic annotation by means of the Alpino parser (van Noord 2006, van Noord et al. 2013) are all suitable corpora for our project. The first two resources provide high-quality data for a limited amount of text, while the last resource provides wide-coverage, but noisy, data. All treebanks follow (with minor modifications) the same annotation standard (van der Wouden et al. 2002a).

2.1 Automatic links

We have investigated the feasibility of generating links to linguistic resources automatically (van der Wouden et al. 2015). As the Taalportaal texts are in XML format, linguistic examples, linguistic terms, etcetera are marked as such and can be "harvested" as such. Example sentences have been selected and translated into queries for corpus tools such as GrETEL (Augustinus et al. 2013), linguistic terms and lexical items that were highlighted, ended up being automatically linked to resources such as an etymological database such as the etymologiebank (van der Sijs 2010) the large (historical) dictionary WNT (De Vries & te Winkel et al. 1864–1998) or to a section in an on-line version of the Dutch reference grammar ANS (Haeseryn et al. 1997).

2.2 Intelligent links

Next to these automatic links, the linguistic data is also enriched with tailor-made links to corpus data (van der Wouden et al. 2015). For this, student assistants with a considerable linguistic schooling have read the linguistic texts, interpreted them and translated their content into queries that address the corpora that seems most fit to them, documenting their choices and considerations.

The web-based corpus query tool PaQu (Odijk 2015, http://zardoz.service.rug.nl:8067/xpath) is our first tool of choice for executing treebank queries. The PaQu interface helps the user to formulate XPATH-queries; it returns matching sentences in the selected corpus, with the option to display the matching nodes in the syntactic dependency graph. It also displays the query being executed along with a brief description. Queries are dynamic, i.e. the user can switch between treebank corpora, or substitute a given lexical item by an alternative. Furthermore, users can select up to three attributes (i.e. lemma, part of speech, dependency relation, etc.) of matching nodes to obtain a frequency distribution of the attribute values. Advanced users can also modify the XPATH query as they see fit. Integration of queries into the electronic version of the SoD will be done by adding links (in the form of an icon) to paragraphs and examples for which queries are available.

(GrETEL (Augustinus et al. 2013, http://nederbooms.ccl.kuleuven.be/eng/gretel) is another a corpus query tool that supports the same XPATH query language as PaQu, but that also provides support for example based query construction, a feature that might be particularly useful to non-expert users.)

2.3 First results

After completion of approx. 1.000 queries that cover the syntactic parts on complementation and modification of adjectives and adpositions, we have learned that creating suitable queries for a given fragment from the SoD requires creativity and careful experimentation, tuning, and documentation (cf. van Engeland & Meertens 2016). Construction of queries is far from deterministic, that is, different annotators have different opinions concerning the most suitable query for a given example or phenomenon. In a surprisingly high number of cases, there are mismatches (in constituent structure, in part-of-speech) between the presentation in the grammar and the treebank annotation. While this makes the development of queries harder, it also underlines the value of the current project: by systematically exploring the way various linguistic examples are annotated in the treebank, we provide a starting point for further corpus exploration for users that have a general linguistic interest but who are not necessarily experts on Dutch treebank annotation.

The manually verified treebanks almost always provide sufficient examples of basic word order patterns for queries that are not restricted to a specific adjective or preposition. For queries that search for a specific lexical head or for less frequent word order patterns, the Lassy Large treebank usually has to be used. In that case, users must be prepared to see also a certain number of false hits. However, there are also examples in the Taalportaal descriptions that cannot be found even in a 700M word corpus. The conclusion that such word orders are not found in the language would be too strong, but it might be a starting point for further research (e.g. does this construction occur only in certain registers or discourse settings?) or for an alternative analysis (e.g. do these cases really involve adjectives?).

3 Extending the Taalportaal: Afrikaans

Only recently, South Afrika has started building a virtual language institute Viva! (http://viva-afrikaans.org/) that aims at developing a digital infrastructure for Afrikaans. Among its goals are study and description of the Afrikaans language and development of comprehensive tools and resources for written and spoken Afrikaans, including digital dictionaries and corpora; language advice is also supplied. Part of the Viva portal is a comprehensive grammar of Afrikaans, which is based on the Taalportaal architecture, and will be part of the Taalportaal infrastructure.

Bibliography

Augustinus, Liesbeth, Vincent Vandeghinste, Inene Schuurman, and Frank van Eynde. 2013. Example-based treebank querying with GrETEL — now also for Spoken Dutch. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (Nodalida 2013)*, 423–428, Oslo, Norway. NEALT Proceedings Series 16.

Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij, and Maarten C. van den Toorn (eds.). 1997. *Algemene Nederlandse Spraakkunst.* Groningen and Deurne: Martinus Nijhoff and Wolters Plantijn. 2nd ed. http://ans.ruhosting.nl/e-ans/index.html.

Landsbergen, Frank, Carole Tiberius, and Roderik Dernison. 2014. Taalportaal: an online grammar of Dutch and Frisian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, & Stelios Piperidis, Reykjavik, Iceland. European Language Resources Association (ELRA).

Van Noord, Gertjan. 2006. At Last Parsing Is Now Operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, ed. by Piet Mertens, Cedrick Fairon, Anne Dister, & Patrick Watrin, 20–42.

Van Noord, Gertjan, Gosse Bouma, Frank van Eynde, Daniel de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In *Essential Speech and Language Technology for Dutch: the STEVIN Programme*, ed. by Peter Spyns & Jan Odijk, 147–164. Springer.

Odijk, Jan. 2015. Linguistic Research with PaQU. *Computational Linguistics in The Netherlands journal* 5, 3–14.

Schuurman, Ineke. 2015. Concept revival: from ISOcat to CLARIN Concept Registry. *CLARIN News* 7 January 2015.

Schuurman, Ineke, Machteld Schouppe, Heleen Hoekstra, and Ton van der Wouden. 2003. CGN, an annotated corpus of spoken Dutch. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, ed. by Anne Abeillé, Silvia Hansen-Schirra, & Hans Uszkoreit, 101–108. Budapest.

Van der Sijs, Nicoline, 2010. Etymologiebank. http://etymologiebank.nl/.

De Vries, Matthias, Lammert A. Te Winkel et al.. (eds.). 1864–1998. *Woordenboek der Nederlandsche taal.* 's-Gravenhage [etc.]: Martinus Nijhoff [etc.]. http://www.inl.nl/.

Van der Wouden, Ton, Gosse Bouma, Matje van de Kamp, Marjo van Koppen, Frank Landsbergen, and Jan Odijk. 2015. Enriching a grammatical database with intelligent links to linguistic resources. Paper delivered at CLARIN 2015, 15–17 October 2015, Wrocław, Poland, accepted for publication in the *Selected Papers of the CLARIN 2015 Conference*.

Van der Wouden, Ton, Heleen Hoekstra, Michael Moortgat, Bram Renmans, and Ineke Schuurman. 2002a. Syntactic Analysis in the Spoken Dutch Corpus (CGN). In *Proceedings of the third International Conference on Language Resources and Evaluation*, ed. by Manuel González Rodríguez & Carmen Paz Suárez Araujo, 768–773. Paris: ELRA.

Van der Wouden, Ton, Heleen Hoekstra, Michael Moortgat,

Ineke Schuurman, and Bram Renmans. 2002b. Syntactische annotatie voor het Corpus Gesproken Nederlands (CGN). *Nederlandse Taalkunde* 7, 335–352.

Van Engeland, Jorik, and Erlinde Meertens. 2016. *Evaluation report pilot links on taalportaal to corpus search interfaces (tpc)*. Technical report, Utrecht University, CLARIN, and University of Groningen.

Stylochronometry and the Periodization of Samuel Beckett's Prose

Dirk van Hulle
dirk.vanhulle@uantwerp.be
University of Antwerp, Belgium

Mike Kestemont
mike.kestemont@uantwerp.be
University of Antwerp, Belgium

Introduction

Probably best known as the author of *En attendant Godot / Waiting for Godot*, Samuel Beckett was not only a bilingual playwright, but also a poet, translator, essayist and novelist. Notably his prose fiction is the focus of this contribution, in which we use quantitative methods to delineate a periodization of Beckett's œuvre. In art studies in general, there is a tradition of distinguishing an 'early' and 'late' period in an artist's work, sometimes with a distinct 'middle' period in between. The late Beethoven sonatas are a good example, or the early Rembrandt's 'smooth' style versus the rough paint surfaces of the late Rembrandt. Nevertheless, it is often difficult to determine exactly when an author's work moves from, say, the 'early' to the 'middle' stage.

In Beckett studies, we find a similar pattern of periodization, ending with the 'late style' (Gontarski, 1997). Peter Boxall (2015: 34) problematizes the idea of periodizing Beckett's œuvre, but admits that it is hard not to parcel it into a beginning, a middle and an end: an early period up to and including the novel *Watt*, written during WOII; a rich middle period up to and including *The Unnamable*; and the later, 'halting' prose after the latter text. Numerous critics have proposed alternative periodizations, resulting in almost as many different periodizations as the number of critics that devised them. We therefore investigated what a non-human 'interpreter' would come up with as a periodization by means of stylometry.

In this paper, we apply methods from stylometry to the problem of periodizing Samuel Beckett's prose. A novelty is that we restrict it to function word frequencies, which are a common object of research in stylometry, but which have hardly been considered in Beckett studies. An exception is Banfield, who suggested a four-phase evolution in Beckett's oeuvre, partially on the basis of linguistic arguments. Our approach is related to 'stylochronometry' (Stamou, 2008), in which a text's writing style is studied as a function of its date of composition, in accordance with recent research into the stylistic development of individual authors, such as Henry James (Hoover, 2014), W. B. Yeats (Forsyth, 1999) or Jack London (Juola, 2007).

Preprocessing

Here, we analyze Beckett's prose fiction, in both French and English. We removed all non-authorial paratexts and only considered lower-case, alphabetical character strings. In the software repository accompanying this contribution (https://github.com/mikekestemont/beckett), we present a tabular overview of the materials we collected, including the publication dates of the editions we used, the text's length, etc. (Note that this repository also holds high-resolution versions of our plots, which will be much more readable on screen.) In terms of chronology, we focus on the moment when Beckett started working on a text in either language (Van Hulle et al., 2015).

We defined a relevant list of function words by extracting an initial list of the 300 most frequent words (MFW) from each corpus. From this list, we have manually removed non-grammatical words, which might correlate too strongly with the topic of particular texts. We refrained from removing common auxiliary verbs or personal pronouns, because they are interestingly tied to a text's narrative perspective, as will become clear below. After this procedure, we were left with 162 function words for the English corpus and 169 for the French, the frequencies of which were scaled using *z*-scores.

Preliminary Analyses

We carried out an exploratory analysis of the material, using principal components analysis (PCA, first two dimensions plotted in Figs. 1-2 for 3,761-word slices). This unsupervised procedure does not yet integrate any chronological information in the analyses, which will help establish whether Beckett's œuvre might have a 'natural' chronological structure with respect to writing style. In Figs. 1-2, the horizontal spread is dominated by a threefold clustering, with some of Beckett's earlier works clustering in the far left. The loadings reveal that these works are characterized by a high frequency of words related to a third-person narrative perspective (*he*, *she*, *his*, *has*, etc.). To the far right, we find a tight sample cloud corresponding to some of Beckett's post-war works, such as *The Unnamable* and *Texts for Nothing*. These texts can be characterized by the use of first-person pronouns (*I*, *me*) in combination with impersonal pronouns such as *it* and *there*, which suggests that these texts focus on the relation between an 'I' and his non-personal surroundings.

Texts from the in-between period (such as *Molloy* and *Watt*) also hold the middle in the horizontal distribution of samples. In both languages, samples from later works jump out with respect to the vertical dimension – e.g. *Worstward Ho* in English, which is characterized by a rich mix of fairly abstract words, with an indeterminate semantics. Both scatterplots horizontally create an opposition between earlier and later writings in Beckett's oeuvre, focusing on an opposition between a first-person and a third-person perspective. In the vertical dimension, both analyses reveal on a vocabulary shift, in Beckett's late writings, towards a more abstract and indeterminate vocabulary.

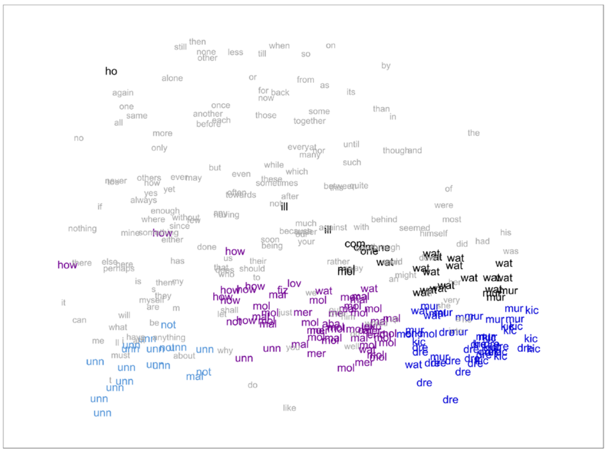


Figure 1: English corpus PCA

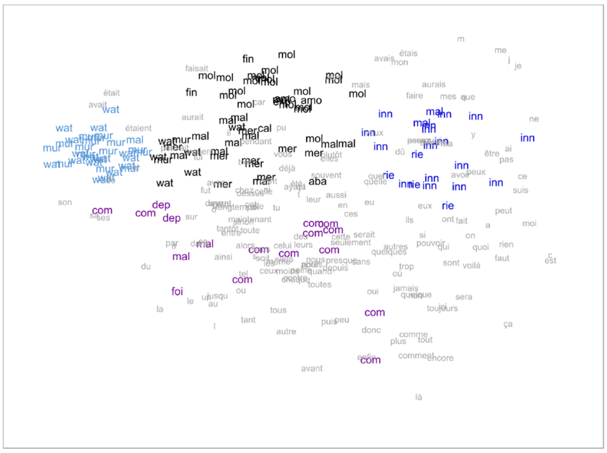


Figure 2: French corpus PCA

Chronological Analyses

The previous analyses were ignorant of the diachronic structure of the data: samples from Beckett's early works could just as easily cluster with later writings. This prevented us so far from identifying clear turning points in Beckett's career. Variability-based Nearest Neighbour Clustering or VNC (Gries et al., 2012) is a method for the diachronic analysis of corpus linguistic data. VNC aims to identify distinct temporal stages, by pinpointing the main turning points. In traditional cluster analyses, each node can be freely combined with any other node in the tree, thus potentially scrambling the original chronological order of the data. VNC adds the constraint that only consecutive nodes, immediately adjacent in time, can form new clusters. This restriction enables analyses in which the chronological structure of the data is reflected in the top branches of trees, representing the main diachronic stages in the data.

We have run VNC on our data (Figures 3-4, Euclidean distances, Ward's linkage), which resulted in clearer insight into the chronological structure of Beckett's œuvre. The English prose, displays a clear initial cluster of Beckett's earliest two novels, *More Pricks than Kicks* and *Dream of Fair to Middling Women*, which lack a French translation. Otherwise, the structures of Figs. 3-4 run remarkably parallel. *Murphy* and *Watt* constitute the second chronological cluster of works, together with the *Nouvelles*. Only at a higher level is the former group paired with the cluster consisting of *Mercier and Camier*, *Molloy* and *Malone Dies*. Interestingly, the original French versions of the *Nouvelles* are joined with the next cluster, whereas the English versions are joined with the previous cluster, which indicates a different status of this collection in both languages. The last major branch for both languages holds the tight clade representing *The Unnamable*, *Texts for Nothing*, *How It Is* and the series of shorter late works. In the English tree, *Worstward Ho* occupies a fairly pronounced position, emphasizing its special status.

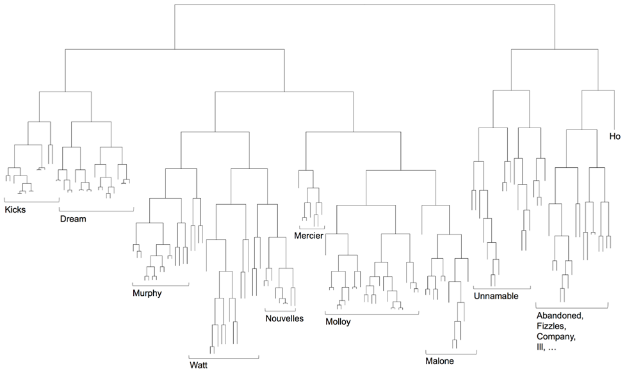


Figure 3: English corpus VNC

The VNC analysis (supplemented by a more complex, bootstrapped analysis, which we do not describe in this abstract) generally supports the periodization of Beckett's oeuvre into an early, middle and late 'cluster'. In English these periods would cover, firstly, Beckett's early works, *More Pricks than Kicks* and *Dream of Fair to Middling*

Women; secondly the mid-career works, ranging from *Murphy* to *Malone dies*; and thirdly, a series of later works starting with *The Unnamable*. From the French prose, a similar periodization arises – although it clearly reflects the absence of any translated counterparts for *More Pricks than Kicks* and *Dream of Fair to Middling Women*.

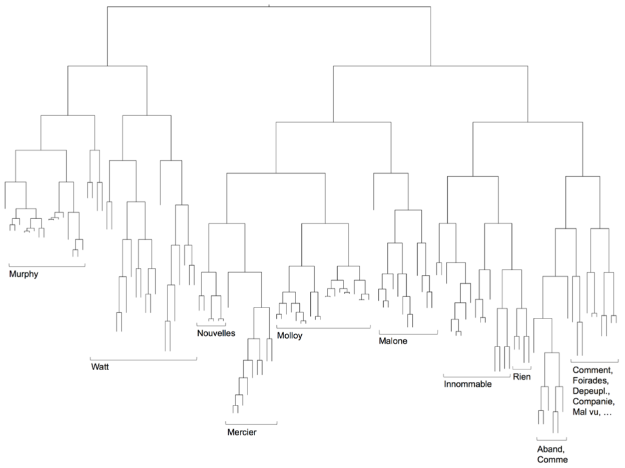


Figure 4: French corpus VNC

Interestingly, our analyses invariably point to the beginning, rather than the end, of *L'Innommable / The Unnamable* as a major stylistic turning point, thus breaking up the unity of the so-called post-war 'trilogy'. Additionally, *From an Abandoned Work / D'un ouvrage abandonné* – notwithstanding its short length – also often emerged as a major watershed. The question, however, is whether this result is to be interpreted as a turning point leading the way for Beckett's later, experimental works, or as a sort of 're-turning point', marking a temporary relapse into the idiom of the *Nouvelles* after *L'Innommable*. This might be a valuable pointer for further research, since this particular text does not seem to have played a significant role in the periodization debate so far. Additionally, it turns out to be more difficult in English than in French to model the transition from a young to a middle Beckett (also in the bootstrapped experiments). This result possibly reflects the fact that the original, English version of *Watt* was written relatively early (during the war), whereas its translation was made much later. In any case, this particular result offers grounds for a re-examination of the difference in evolution between Beckett's French and English prose production, and in particular the role of *Watt* as a transitional novel.

Bibliography

Banfield, A. (2003). Beckett's Tattered Syntax, *Representations*, **84**: 6-29.

Boxall, P. (2015). Still Stirrings: Beckett's Prose from Texts for Nothing to Stirrings Still. In Van Hulle, D. (ed), *The New Cambridge Companion to Samuel Beckett*. Cambridge: Cambridge University Press, pp. 33-47.

Gontarski, S. E. (1997). Staging Himself, or Beckett's Late Style in the Theatre. *Samuel Beckett Today / Aujourd'hui*, **6**: 87-97.

Forsyth, R. S. (1999). Stylochronometry with substrings, or: a poet yound and old. *Literary and Linguistic Computing*, **14**: 467-78.

Gries, S. T. and Hilpert, M. (2012). Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In Nevalainen, T. and Traugott, E. C. (eds), *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, pp. 134-44.

Hoover, D. L. (2014). A Conversation Among Himselves. Change and the Styles of Henry James. In Hoover, D. L., Culpeper, J. and O'Halloran, K. (eds), *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. New York: Routledge, pp. 90-119.

Juola, P. (2007). Becoming Jack London. *Journal of Quantitative Linguistics*, **14**: 145-47.

Stamou, C. (2008). Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, **23**: 181-99.

Van Hulle, D. and Verhulst, P. (2015). Introduction: A Beckett Continuum and a Chronology of Beckett's Writings. In Van Hulle, D. (ed), *The New Cambridge Companion to Samuel Beckett*. Cambridge: Cambridge University Press, pp. XVII-XXXII.

Van Hulle, D. and Kestemont, M. (2016). Periodizing Samuel Beckett's works: a Stylochronometric Approach. *Style* 50: forthcoming.

Qu'est-ce qu'un texte numérique? A New Rationale for the Digital Representation of Text

Joris J. Van Zundert
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands - Royal Netherlands Academy of Arts and Sciences

Tara L. Andrews
tara.andrews@kps.unibe.ch
Universität Bern, Switzerland

Over half a century ago Briet (1951) famously asked "Qu'est-ce que la documentation?" She proposed a definition much more fluid than the scraps of writing on paper we usually associate with the term: any object, even an uninscribed stone, becomes a document as soon as it is used to communicate some fact (e.g. in a museum's geological collection). In a similar vein we argue in this paper for a renewed consideration of the nature of text in the digital realm, or rather the nature that has been imposed upon it both by information-technological and methodological constraints.

Text is by its nature both discrete and continuous. The glyphs of a writing system are combined into words, which are arranged into phrases, sentences, paragraphs, quotations, and so on; these discrete formations can be arbitrarily complex, but they are communicated through the atom of the glyph. On the other hand, very many elements of the meaning of a text—e.g. narratological elements such as the theme of isolation in *Do Androids Dream of Electric Sheep?*, intertextual references and referents (Miola 2000), the concept of *writerly text* (Barthes, 1974), or even the visual presentations of individual glyphs, defy discrete boundaries. Yet the technologies we have used overwhelmingly for the large-scale production of text—from the printing press to signal transmissions and thence to digital models—treat text as a code of discrete characters; moreover, once text moves from "print" to "signal" (Petzold, 2000) it becomes specifically a single stream of these discrete characters. This can already present some problems to scholars who wish to remediate texts not coded by machine, e.g. manuscripts, to the digital domain for humanistic enquiry. To date scholars who encounter these problems generally consider them an acceptable trade-off for the benefits of access, exchange, and computational tractability that the digital medium offers.

If we want our digital model of a text to extend beyond the semiotic registration of a single stream of characters, however, problems emerge. McGann (2004) summarizes this well: "Print and manuscript technology represent efforts to mark natural language so that it can be preserved and transmitted. It is a technology that constrains the shapeshiftings of language, which is itself a special-purpose system for coding human communication. Exactly the same can be said of electronic encoding systems. In each case constraints are installed in order to facilitate operations that would otherwise be difficult or impossible." Everyone who has ever worked on the remediation of a physically-inscribed text into the digital medium has implicitly conceptualized it at some point as a single stream of discrete characters, simply because this is how we have always known text, ultimately, to be represented digitally. Transcription is thus exactly the sort of constraint to which McGann refers: there is a single stream of text, and only once it exists can it be subdivided and classified as having certain descriptive, structural, or logical characteristics, or even relationships to other portions of text.

The model of text as a single stream of discrete characters also informs the concept of XML markup.[1] As Pierazzo (2011) notes, "The editor will first transcribe a primary source, thereby creating a transcription; this transcription will be corrected, proofread, annotated, and then prepared for publication." The predominant model for the process of annotation to which Pierazzo refers is TEI-XML, that is, embedded markup within the hierarchical structure required by XML. Conceptually a tree structure expressed in XML can be perfectly reorderable, and so not bound to a single valid serialization. With text encoding, however, the ordering constraint remains and must be preserved, however marked up it is and subdivided into branches. Standoff markup schemes similarly assert a single stream—each scheme relies on the stability of the range notation it uses for the underlying text stream—but the markup itself is reorderable. This reorderable property of standoff markup takes us a small step closer to our goal: to recast and represent digital text as something more than a single stream of discrete tokens.

Standoff markup essentially applies a graph model. This is the feature of standoff that makes it so valuable to its proponents—in a graph, unlike a tree, overlapping ranges of markup are easily handled. Some of these standoff markup elements might even include alternative readings to a given text range, or a reference to a different text that elucidates the contextual meaning of the range in question. From there it is only a small step to the conception of a text that is itself more complex than a single stream. Such conceptions emerge already with the different implementations of variant graphs for collations of "multi-version documents" (Schmidt and Colomb, 2009; see also Dekker et al., 2015; Andrews and Macé, 2013; Jänicke et al., 2014). While so far the variant graph has been used only for word-by-word comparisons of text versions, the concept can be extended radically farther. We will demonstrate this, building on prior work in this direction (e.g. Marcoux, Sperberg-McQueen, and Huitfeldt, 2013). We will show in our paper, for instance, how a graph-based digital text can model collections of stories preserved in manuscripts, without giving primacy to any one ordering of the story sequence; that is, it can contain both the collated text of each story and the variety of story order within the tradition. Likewise the same narrative element, inserted into multiple stories, can be represented as such.

The representation of information as entities with relationships between them—in other words: as a graph—is not a new idea. RDF works exactly this way, and the theoretical ideas behind hypertext are not dissimilar (cf. Nelson, 1993). The ease with which complex graphs can be modelled in software, however, has massively increased in the last five years. Moreover, it is now reasonably straightforward to create an initial graph representation from a TEI-encoded XML file, and nearly as straightforward to produce multiple "flavors" of TEI XML from a single graph—for example, a document-oriented representation alongside a logical-text-oriented representation, such as those created in the Faust edition project (Brüning et al., 2012).

Graph representation does not relieve us from the constraint that digital text must be formed from discrete atomic units (characters) and discrete compound digital units. What it does allow us to do, however, is to digitally model the shades of meaning, ambiguity, and uncertainty of text—the aspects of a text's meaning or interpretation that are not by their nature discrete at all. So far, uncertainty of interpretation has been all but explicitly avoided in the scholarly digital space. McGann (2004) observes: "In the case of a system like the Text Encoding Initiative (TEI), the system is designed to 'disambiguate' entirely the materials to be encoded." The 1700+ pages of the TEI guidelines bear witness to the difficulty of this task; element names and their intended usages are defined as precisely as possible, with illustrative examples.

The use of tags and attributes in a TEI-encoded file is itself an act of interpretation—it is precisely in the labelling of the elements of an encoding and of the relationships between those elements that much of the scholar's interpretation of the encoded text lies. This is equally true of a graph encoding. The prolixity of the Guidelines is in fact a result of the attempt to constrain the scholar's interpretation of a given tag to a disambiguated, and thereby discrete, set of meanings. Rather than shrinking from uncertainty of meaning, however, rather than resolving it in an occasionally misleading or (paradoxically) unhelpful attempt to ease computational analysis, how much better if the uncertainty can be retained in the model, in a manner that is nonetheless machine-parseable? Interpretation remains inherent in the labels that are chosen for the properties in a graph model; these may be taken from the TEI, or they may reflect another interest of the scholar. Either way, a graph-based text model by its nature includes computationally tractable information about how these properties relate to each other.

In representing "fuzziness" or ambiguity of interpretation, we can perhaps follow the lead of those who have struggled with similar problems of representation of time. If text can be represented as a graph, mostly-sequential but with scope for fluidity, then interpretative elements—say, picking out the theme of isolation within the text—can be represented as spanning a continuous or disjoint range of text sequences, beginning not before word A but certainly by word B, and continuing at least as far as word C but certainly not beyond word D. And so on. If scholars disagree on the correct identification of a person in a historical text, both hypotheses can be represented without innate primacy being given to one over the other, because a graph does not enforce a single ordering of its elements.

The text model we are advocating, then, amounts to a form of knowledge representation; in representing "the text", we are actually representing a collection of our knowledge, understanding, and interpretation of that text, in a form that can be analyzed and processed by a machine. The more we can encode texts as computable knowledge, the greater the power computational methods will have in textual scholarship of all forms.

Bibliography

Andrews, T.L., Macé, C., 2013. "Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata." *Literary and Linguistic Computing* 28 (4): 504–521.

Barthes, Roland.*S/Z*. Trans. Richard Miller. London: Cape, 1975.

Briet, S., 1951.*Qu'est-ce que la Documentation?*, Paris: Édit. Translated as S. Briet, *What is Documentation?: English Translation of the Classic French Text*, Lanham, Md: Scarecrow Press. Available at: http://ella.slis.indiana.edu/~roday/briet.htm [Accessed 1 November 2015].

Brüning, G., et al., 2012. "On the dual nature of written texts and its implications for the encoding of genetic manuscripts". Presented at the Digital Humanities 2012, Hamburg.

Dekker, R.H., van Hulle, D., Middell, G., Neyt, V., van Zundert, J., 2014. "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project." *Digital Scholarship in the Humanities* 30 (3): 452–470.

Goldfarb, C.F., The Roots of SGML – A Personal Recollection. Available at: http://www.sgmlsource.com/history/roots.htm [Accessed 1 November 2015].

Grossner, K. and Meeks, E., 2014. "Topotime: Representing Historical Temporality." Presented at the Digital Humanities 2014, Lausanne .

Jänicke, S., Geßner, A., Büchler, M., Scheuermann, G., 2014. "5 Design Rules for Visualizing Text Variant Graphs." Presented at the Digital Humanities 2014, Lausanne.

Marcoux, Y., Sperberg-McQueen, M.C. & Huitfeldt, C., 2013. Modeling overlapping structures: Graphs and serializability. In *Proceedings of Balisage: The Markup Conference 2013*. Balisage Series on Markup Technologies. Balisage: The Markup Conference 2013. Montréal, Canada. Available at: http://www.balisage.net/Proceedings/vol10/html/Marcoux01/BalisageVol10-Marcoux01.html [Accessed 1 November, 2015].

McGann, J., 2004. Marking Texts of Many Dimensions. In S. Schreibman, R. Siemens, & J. Unsworth, eds. *A Companion to Digital Humanities*. Oxford: Blackwell. Available as chapter 16 at: http://www.digitalhumanities.org/companion/ [Accessed 1 November 2015].

Miola, R.S., 2000. Seven Types of Intertextuality. In M. Marrapodi, ed. *Shakespeare, Italy, and Intertextuality*. Rome: Bulzoni Editore, pp. 23–38.

Nelson, T.H., 1993. *Literary Machines. The report on, and of, Project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom.* First published 1981., Mindfull Press.

Petzold, C., 2000.*Code: The Hidden Language of Computer Hardware and Software*. Redmond: Microsoft Press.

Pierazzo, E., 2011. "A rationale of digital documentary editions". *Literary and Linguistic Computing* 26 (4): 463–77.

Renear, A., 2004. Text Encoding. In S. Schreibman, R. Siemens, & J. Unsworth, eds. *A Companion to Digital Humanities*. Oxford: Blackwell. Available as chapter 17 at: http://www.digitalhumanities.org/companion/ [Accessed 1 November 2015].

Schmidt, D., Colomb, R., 2009. "A data structure for representing multi-version texts online." *International Journal of Human-Computer Studies* 67: 497–514.

Notes

[1] While this is not stated explicitly anywhere of which we are aware, the underlying assumption is plain in, e.g., Goldfarb's discussion of the roots of SGML (1996) or Renear's definition of text encoding (2004).

Music notation addressability

Raffaele Viglianti
rviglian@umd.edu
University of Maryland

Introduction

How can one virtually 'circle' some music notation as one would on a printed score? How can a machine interpret this 'circling' to select and retrieve the relevant music notation in digital format? This paper will introduce the concept of addressability for music notation, on the basis of a comparison with textual addressability as defined by Michael Witmore (2010). Additionally, the paper will report on the work of *Enhancing Music notation Addressability* (EMA), a NEH-funded one-year project that has developed methods for addressing arbitrary portions of encoded music notation on the web.

Many Digital Humanities projects are concerned with the digitization of cultural objects for varied purposes of study and dissemination. Theorists such as Willard McCarty (2005) and Julia Flanders (2009) have highlighted the fact that digitization involves the creation of a data model of a cultural object, whereby scholarly interpretation and analysis is inevitably included in the model. Editorial projects in literary studies, for example, often model sources by encoding transcription and editorial intervention with the Text Encoding Initiative (TEI) format. The ability to identify and name textual structures is a fundamental operation in the creation of such models. Michael Witmore has called text a "massively addressable object" (2010); that is, given certain abstractions and conventions, it is possible to identify areas of a text such as characters, words, as well as chapters or proper names. Reading practices influence and contribute to the development of such conventions and abstractions, but, Witmore argues, addressability is a textual condition regardless of technology. With digital texts, modes of address become more abstract, so that arbitrary taxonomies can be identified as well as more established ones. To exemplify a more abstract mode of address, Witmore suggests items "identified as a 'History' by Heminges and Condell in the First Folio". This enhanced addressability available in a digital context is the engine for textual analysis and scholarly discourse about digital text.

This idea of addressability is arguably applicable to many more kinds of "text", including music notation; indeed, addressing units of music notation (such as measures, notes, and phrases) has long been a powerful instrument in musicology for both analysis and historical narrative.[1] Music notation, however, is more complicated to represent digitally than text. Human-computer interaction has since its early days been built around the concept of character and line, which makes dealing with "plain" text a fairly straightforward matter for many basic operations; counting the number of characters in a given plain text document is trivial in any digital environment.[2] Music notation, on the other hand, requires substantial computational modelling even for the simplest musical text before any further operation is possible. This is particularly evident when music notation is represented with markup, which implies a system based on characters and lines. There are many different ways of representing a single note; some aspects are common to all representation systems, such as information about pitch and duration, but some systems will prioritize certain aspects over others. To give a simple example, one system may represent beams (ligatures between flagged notes, usually shorter in duration), while others may ignore them altogether.[3]

Nonetheless, there are simple units that are typically represented by all music notation systems for common western music notation, such as measure, staff (or instrument), and beat. The EMA project, therefore, developed a URI scheme and an Application Programming Interface (API) to make it possible to target music notation resources on the web regardless of their format. Such a scheme may facilitate (and in some cases enable) a number of activities around music notation documents published on the web. The following table gives a few basic examples of how an implementation of the URI scheme could be useful to musicological research:

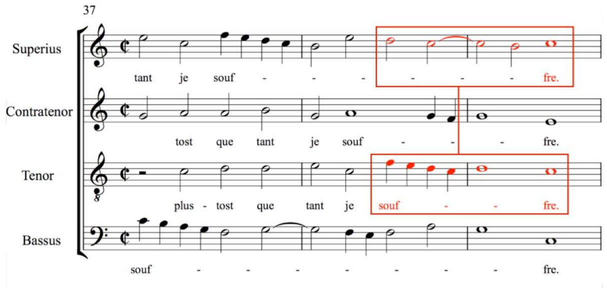
Scholarly	Visual	Procedural
Analysis: being able to address components of music notation for analytical purposes. Example: precisely identify start and end of a pedal tone in Bach's Prelude no. 6 in D Minor, BWV 851.	Rendering: rendering music notation in an interactive environment such as a browser or a tablet requires the ability to cut up a large music document. For example to show only the number of measures that fit in a given space.	Processing: extracted portions of music notation can be passed on to another process. For example, given the MEI encoding of the Overture to Mozart's Don Giovanni, extract the string instrument parts and send them to another program that will return an harmonic analysis.
Citation: quote a passage from an encoded music notation file. For example the timpani in the opening bars of the Overture to Mozart's Don Giovanni.	Highlighting: address a segment of music notation to highlight it in a visual context (e.g. with color).	

The EMA project has particularly focused on facilitating citation and attribution of credit, as is discussed in the "Evaluation" section below.

A brief overview of the specification

The specification was created to provide a web-friendly mechanism for addressing specific portions of music notation in digital format. This is not unlike the APIs often provided by image servers for retrieving specific portions of an image. Such servers typically operate on a given large image file and are able to return different zoom levels and coordinate spaces. The International Image Interoperability Framework (IIIF) has recently created an API to generalize interaction with image providers, so that it can be implemented across multiple servers and digital libraries. IIIF was used as a model for the Music Addressability API created for EMA and briefly described here.

Consider the following example,[4] and the notation highlighted in the boxes:


37
Superius
tant je souf - - - - - - - fre.
Contratenor
tost que tant je souf - - - fre.
Tenor
plus - tost que tant je souf - - fre.
Bassus
souf - - - - - - - - fre.

The highlighted notation occurs between measure 38 and 39, on the first and third staves (labelled *Superius* and *Tenor* — this is a renaissance choral piece). Measure 38, however, is not considered in full, but only starting from the third beat. This selection can be expressed according to a URI syntax:

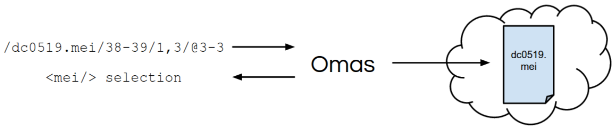
/{identifier}/{measures}/{staves}/{beats}/
/dc0519.mei/38-39/1,3/@3-3

The measure is expressed as a range (38-39), staves can be selected through a range or separately with a comma (1,3), and the beats are always relative to their measure, so @3-3 means the third beat of the starting measure to the third beat of the ending measure.[5] In this specification the beat is the primary driver of the selection: it allows for precise addressability of contiguous as well as non-contiguous areas.

Music notation, however, occasionally breaks rules in favor of flexibility. Cadenzas, for example, are ornamental passages of an improvisational nature that can be written out with notation that disregards a measure's beat, making it impossible to address subsets of the cadenza wit the syntax discussed above. While EMA's URI scheme offers the granularity sufficient to address the vast majority of western music notation, a necessary future improvement on the API is, indeed, an extension that would make it possible to address music notation with more flexible beat.

Evaluation

In order to evaluate the specification, EMA has created an implementation of the API as a web service. While the URI specification can be absolute from a specific representation, the implementation must know how to operate on specific formats. The web service that we coded operates on the The Music Encoding Initiative format and is called Open MEI Addressability Service (Omas).[6] Omas interprets a conformant URI, retrieves the specified MEI resource, applies the selection, and returns it. An additional parameter on the URI can be used to determine how "complete" the retrieved selection should be (whether it should, for example, include time and key signatures, etc.).



Similarly to an image server, Omas assumes that the information specified by the URL can be retrieved in the target MEI file. If requested, the web service can return metadata information about an MEI file, such as number of measures, staves, beats and their changes throughout the document. This can be used to facilitate the creation of URL requests able to return the selection required.

Finally, EMA partnered with the *Du Chemin: Lost Voices* project to model a number of micro-analyses addressing music notation from their existing collection of MEI documents. In a second phase of the project, the analyses have been re-modeled as Linked Open Data according to the Nanopublication guidelines.[7] Each EMA nanopublication addresses an arbitrary portion of music notation using the URL specification described here. Omas operates as a web service to connect the nanopublications with the collection of MEI files in *Du Chemin*.

Bibliography

Babbit, M. (1965). The use of computers in musicological research. *Perspectives of New Music*, 3(2): pp. 74–83.

Flanders, J. (2009). Data and Wisdom: Electronic Editing and the Quantification of Knowledge. *Literary and Linguistic Computing*, 24(1): pp. 53–62.

McCarty, W. (2005). Chapter 1 - Modelling. *Humanities Computing*, London: Palgrave Macmillan.

Witmore, M. (2010). *Text: A Massively Addressable Object*. http://winedarksea.org/?p=926.

Notes

[1] When talking about music in general, it is important to say that addressing written music notation is not the only instru-

ment of the musicologist. Music exists on several domains besides the written or "graphemic" one, each addressable in its own way (see Babbitt 1965). For the purpose of this paper, we focus on written Western music notation, because it shares features with written language and for its prominent role in musicological discourse.

[2] Modern computing systems are able to support complex ancient and modern writing systems, including those requiring right-to-left strings and compound symbols. The Unicode Consortium has been at the forefront of the internationalization of computing systems. Nonetheless, computationally speaking, a "string" of text remains a sequence of characters even in more complex representations. Indeed, many compound Unicode characters still retain sequentiality, i.e. one component comes after the other and the compound symbol only makes sense if they are in the correct order. Music notation is not a string of text; therefore this is not possible.

[3] By grouping notes together, beams provide important—but somewhat secondary to pitch and duration—information to the reader of a music score, such as a performer, a musicologist, or an algorithm.

[4] Taken from *Du Chemin: Lost Voices* project, at http://digitalduchemin.org.

[5] A complete description of the URI scheme and the API is available at: https://github.com/umd-mith/ema/blob/master/docs/api.md.

[6] A demo is available at http://mith.us/ema/omas/.

[7] Nanopublication is an ontology for publishing scientific data: http://nanopub.org. The Nanopublication server for *Du Chemin: Lost Voices* is available at: http://digitalduchemin.org/np/.

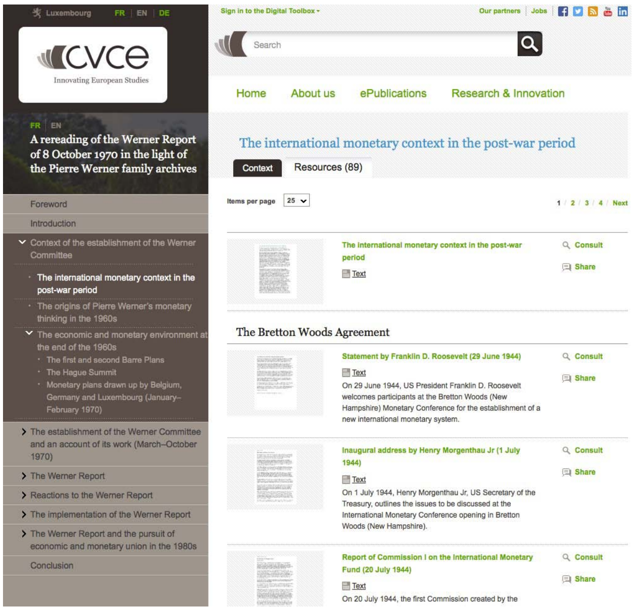
Introducing HistoGraph 2: Exploration of Cultural Heritage Documents Based on Co-Occurrence Graphs

Lars Wieneke
lars.wieneke@cvce.eu
CVCE, Luxembourg

Marten Düring
m.duering@zoho.com
CVCE, Luxembourg

Daniele Guido
daniele.guido@cvce.eu
CVCE, Luxembourg

In this paper we discuss our strategy for the creation and exploration of graphs based on co-occurrences of named entities using the recently developed open source version of histograph and present a live demo. histograph builds on previous work conducted in cooperation with several European partners in the FP7 funded project CUbRIK but has been further developed by the Centre Virtuel de la Connaissance sur l'Europe (CVCE) in 2015. It combines the graph-based exploration of large cultural heritage collections with crowd-based indexation. histograph opens up a new perspective on CVCE's collections which contain some 20.000 digitized text documents, photos, audio recordings and videos that document the history of European integration since 1945.



FR | EN
A rereading of the Werner Report of 8 October 1970 in the light of the Pierre Werner family archives

- Foreword
- Introduction
- Context of the establishment of the Werner Committee
  - The international monetary context in the post-war period
  - The origins of Pierre Werner's monetary thinking in the 1960s
  - The economic and monetary environment at the end of the 1960s
    - The first and second Barre Plans
    - The Hague Summit
    - Monetary plans drawn up by Belgium, Germany and Luxembourg (January–February 1970)
- The establishment of the Werner Committee and an account of its work (March–October 1970)
- The Werner Report
- Reactions to the Werner Report
- The implementation of the Werner Report
- The Werner Report and the pursuit of economic and monetary union in the 1980s
- Conclusion

The international monetary context in the post-war period

Context | Resources (89)

Items per page 25 | 1 / 2 / 3 / 4 / Next

The international monetary context in the post-war period
Text
Consult
Share

The Bretton Woods Agreement

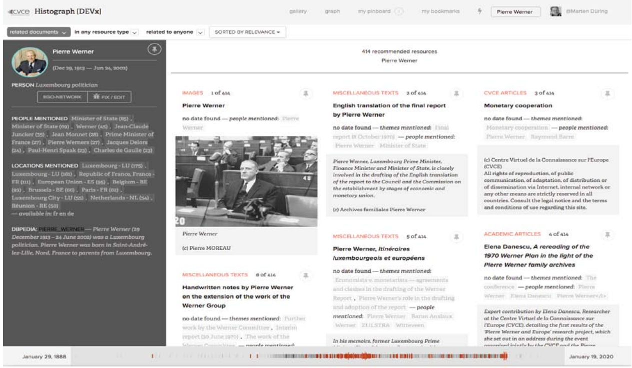
Statement by Franklin D. Roosevelt (29 June 1944)
Text
On 29 June 1944, US President Franklin D. Roosevelt welcomes participants at the Bretton Woods (New Hampshire) Monetary Conference for the establishment of a new international monetary system.
Consult
Share

Inaugural address by Henry Morgenthau Jr (1 July 1944)
Text
On 1 July 1944, Henry Morgenthau Jr, US Secretary of the Treasury, outlines the issues to be discussed at the International Monetary Conference opening in Bretton Woods (New Hampshire).
Consult
Share

Report of Commission I on the International Monetary Fund (20 July 1944)
Text
On 20 July 1944, the first Commission created by the
Consult
Share

Figure 1: A screenshot from a CVCE ePublication on the Werner report. The navigation on the left shows the hierarchically organized themes, on the right a list of expert-curated documents

histograph adds an explorative approach to the hierarchically organized, expert-curated collections of CVCE.eu: Users decide what interests them and find their own path through the collections. A user who is interested in Pierre Werner will for example see a page similar to the screenshot below.



Pierre Werner
(Dec 29, 1913 — Jun 24, 2002)
PERSON *Luxembourg politician*

PEOPLE MENTIONED Minister of State (85), Minister of State (69), Werner (45), Jean-Claude Juncker (35), Jean Monnet (28), Prime Minister of France (27), Pierre Werners (27), Jacques Delors (24), Paul-Henri Spaak (23), Charles de Gaulle (23)

LOCATIONS MENTIONED Luxembourg - LU (175), Luxembourg - LU (161), Republic of France, France - FR (111), European Union - ES (95), Belgium - BE (93), Brussels - BE (66), Paris - FR (62), Luxembourg City - LU (55), Netherlands - NL (54), Réunion - RE (50)
— *available in:* fr en de

DBPEDIA: PIERRE_WERNER — *Pierre Werner (29 December 1913 – 24 June 2002) was a Luxembourg politician. Pierre Werner was born in Saint-André-lez-Lille, Nord, France to parents from Luxembourg.*

414 recommended resources
Pierre Werner

IMAGES 1 of 414
Pierre Werner
no date found — *people mentioned:* Pierre Werner
Pierre Werner
(c) Pierre MOREAU

MISCELLANEOUS TEXTS 2 of 414
English translation of the final report by Pierre Werner
no date found — *themes mentioned:* Final report (8 October 1970) — *people mentioned:* Pierre Werner Minister of State
Pierre Werner, Luxembourg Prime Minister, Finance Minister and Minister of State, is closely involved in the drafting of the English translation of the report to the Council and the Commission on the establishment by stages of economic and monetary union.
(c) Archives familiales Pierre Werner

CVCE ARTICLES 3 of 414
Monetary cooperation
no date found — *themes mentioned:* Monetary cooperation — *people mentioned:* Pierre Werner Raymond Barre
(c) Centre Virtuel de la Connaissance sur l'Europe (CVCE)
All rights of reproduction, of public communication, of adaptation, of distribution or of dissemination via Internet, internal network or any other means are strictly reserved in all countries. Consult the legal notice and the terms and conditions of use regarding this site.

MISCELLANEOUS TEXTS 6 of 414
Handwritten notes by Pierre Werner on the extension of the work of the Werner Group
no date found — *themes mentioned:* Further work by the Werner Committee, Interim report (30 June 1970), The work of the Werner Committee — people mentioned:

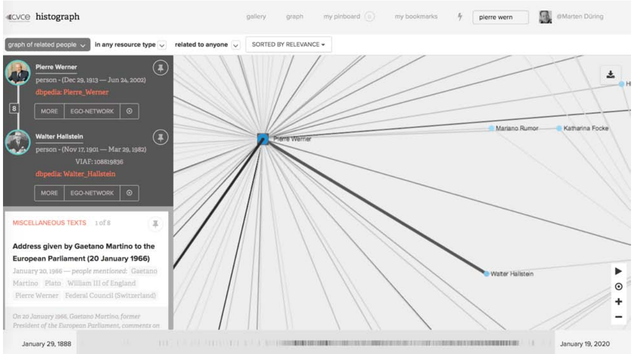
MISCELLANEOUS TEXTS 5 of 414
Pierre Werner, *Itinéraires luxembourgeois et européens*
no date found — *themes mentioned:* Economists v. monetarists — agreements and clashes in the drafting of the Werner Report, Pierre Werner's role in the drafting and adoption of the report — *people mentioned:* Pierre Werner Baron Ansiaux Werner ZIJLSTRA Witteveen
In his memoirs, former Luxembourg Prime

ACADEMIC ARTICLES 4 of 414
Elena Danescu, *A rereading of the 1970 Werner Plan in the light of the Pierre Werner family archives*
no date found — *themes mentioned:* The conference — *people mentioned:* Pierre Werner Elena Danescu Pierre Werner</i>
Expert contribution by Elena Danescu, Researcher at the Centre Virtuel de la Connaissance sur l'Europe (CVCE), detailing the first results of the 'Pierre Werner and Europe' research project, which she set out in an address during the event organised jointly by the CVCE and the Pierre

January 29, 1888
January 19, 2020

Figure 2: A search for „Pierre Werner" reveals a short biographical overview with a list of frequently co-occurring entities (left column) and a gallery view of related documents which can be filtered e.g. by resource type and time

The left column provides an overview of Pierre Werner's biography and a list of persons with whom he co-occurs in the document base. The middle column lists all documents in which he is mentioned. The graph view provides a bird's-eye-perspective on the people who co-occur with Werner. The graph is interactive and reveals the documents which constitute a co-occurrence relationship.



Pierre Werner
person - (Dec 29, 1913 — Jun 24, 2002)
dbpedia: Pierre_Werner

Walter Hallstein
person - (Nov 17, 1901 — Mar 29, 1982)
VIAF: 108819836
dbpedia: Walter_Hallstein

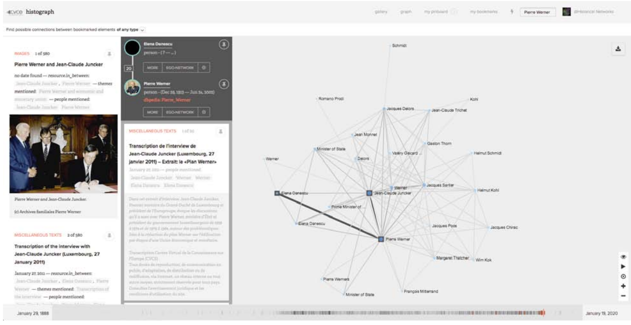
MISCELLANEOUS TEXTS 1 of 8

Address given by Gaetano Martino to the European Parliament (20 January 1966)
January 20, 1966 — people mentioned: Gaetano Martino Plato William III of England Pierre Werner Federal Council (Switzerland)

On 20 January 1966, Gaetano Martino, former President of the European Parliament, comments on

Figure 3: A graph representation shows with whom Pierre Werner co-occurs in documents. A click on a link lists the documents which mention two entities, here Werner and Walter Hallstein

histograph uses a Neo4j (Neo4j Console, 2016) graph database to store relations between entities. This approach facilitates queries that would be computationally expensive in relational databases but are easily available in graph databases, such as the calculation of paths between entities that are not directly connected. Figure 4 shows the result of a query for all documents, which connect three persons.



Find possible connections between bookmarked elements of any type

IMAGES 1 of 580

Pierre Werner and Jean-Claude Juncker
no date found — resource.in_between:

Pierre Werner and Jean-Claude Juncker.
(c) Archives familiales Pierre Werner

MISCELLANEOUS TEXTS 2 of 580

Transcription of the interview with Jean-Claude Juncker (Luxembourg, 27 January 2011)
January 27, 2011 — resource.in_between:

Elena Danescu
person - (? — ...)

Pierre Werner
person - (Dec 29, 1913 — Jun 24, 2002)
dbpedia: Pierre_Werner

MISCELLANEOUS TEXTS

Transcription de l'interview de Jean-Claude Juncker (Luxembourg, 27 janvier 2011) – Extrait: le «Plan Werner»

Figure 4: A query for paths between Pierre Werner, Elena Danescu and Jean-Claude Juncker reveals all relevant documents as well as an interactive graph of co-occurrences (who co-occurs with whom)

In contrast to the hierarchical, expert-curated collections, histograph enables an interest driven exploration by the users and provides them with an effective way to retrieve and explore the relations which are of interest to them. Compared with the museum-like order of the classical CVCE collections, histograph models a more or less targeted visit to an archive, which holds the promise of serendipitous discoveries.

Methods for the annotation of named entities such as persons, institutions and places have reached a very high degree of maturity and are used in different applications. For histograph we tested a variety of web services for the detection and disambiguation of entities (NER) among them TextRazor (TextRazor - The Natural Language Processing API, n.d.) and YAGO (Max-Planck-Institut für Informatik: YAGO, 2016). While these services perform well depending on the context of use (language, domain, etc.) even in a best case scenario they only allow to identify the occurrence of an entity within a logical unit such as a text but do not allow to clarify the nature of the connection between persons, places and organisations that occur together in the same unit. In our previous experiences with the detection and identification of faces in historical image sets (Wieneke et al., 2014) the format and context of the images as official photographs of specific events enabled us to understand the nature of the relationship as more easily defined: the simple working hypothesis that persons that co-occur together in a photo have some kind of connection proved to be very efficient. In the case of text however, these semantics are significantly more complex.

We therefore decided to follow two directions: first based on the nature of the text document, where letters for example constitute a relationship between the sender and the receiver and second through a mathematical modelling of the relationship based on the distance between entities in the text and their distribution within the corpus. More specifically, we work with Jaccard distances and co-occurrence frequencies weighed by tf-idf specificity. The latter step became necessary as not all of our text documents fall in the category of clearly structured formats such as letters and even if they do, a strictly format based co-occurrence approach could hide interesting relationships that would foster exploration, e.g. in our case a written exchange between two politicians where they discuss a third person.

Entities who appear within a certain distance from each other can be linked based on the assumption that their co-occurrence in the text is not arbitrary and that there is a high probability that they have *something* to do with each other. The boundaries for such relations can be defined freely, for example based on sentences, paragraphs, or documents. Alternatively, a window approach considers all entities related if they appear within *n* characters from each other. In addition, such co-occurrences can be further defined by the recognition of the semantic relationships within a sentence. We have, however, not yet systematically tested their performance and therefore limit ourselves here to a more basic approach.

Despite all efforts to further specify the significance of such relations, co-occurrences are elusive:

- It is hard to further specify what more or less connected mean
- It is hard to further specify which types of relations are at play
- It is hard to assess which relations were missed because entities were named differently

A graph representation of such data can only allow rather general statements: We can reasonably assume that entities, which co-occur often, are more connected than those who do not. We can also assume that entities, which never co-occur together, are less or not at all connected. Finally, we can assume that entities, which tend to cluster together, are more likely to have something to do with each other without further specifying their relationship.

In contrast, most social network analysis (SNA) research questions require very well defined relations given that they treat social networks as models of highly complex social interactions. Here, a graph represents a meaningful reduction of such complexity and allow insights into specific dimensions of social relations. Graphs are used to represent and/or illustrate highly complex matters, which are otherwise hard to express. Such data is typically generated and curated for the purpose of specific research questions and its value is limited to their respective context in which it was created.

CVCE and cultural heritage institutions in general however address heterogeneous user groups and wish to make available their data to different audiences including educators, researchers and interested laymen. A graph representation of relations between entities therefore will serve different purposes than in a SNA context. Here, graphs need not be meaningful models of social relations but need to be multi-purpose means for the discovery of materials and acknowledge wildly different interests. This means that it is impossible to predict, which relations a user will consider relevant.

Against this background we embrace co-occurrences despite their inherent shortcomings. In order to improve the quality of the various relations we display we need to balance a) higher quality or more meaningful relations and b) the potential to make highly unexpected, yet meaningful discoveries in the data. In order to achieve this, we decided against a rigid ontology of relationship types, which would significantly limit the chances for unexpected discoveries, beyond format based assumptions. Instead we use text synopses and full document text for the generation of relations and filters on time periods, media types and entities in order to further specify our graphs. In addition, we use a two-fold crowd-based approach: generic crowds help clean the data and report obvious mistakes; expert crowds provide high-quality annotations which required a highly domain-specific knowledge. Generic crowds perform tasks which do not require specialist knowledge, for example an answer to the question “Is ‘N.A.T.O.’ a person?” Expert crowds are identified based on past performance and are presented with tasks related to entities and documents they have worked on before.

The resulting graphs can be considered hybrids: They are inasmuch based on co-occurrences as they are on user specifications. Such graphs, we argue, have the potential to meet both the necessities for the automated generation of graphs and still provide meaningful structural information which can be the starting point for a deeper investigation of the materials. We are committed to further increase the quality of co-occurrence relationships in the future.

Bibliography

Max-Planck-Institut für Informatik: YAGO (2011). https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/ (accessed 3.01.2016).

Neo4j Console (2016). http://console.neo4j.org/ (accessed 2.06.2015).

TextRazor - The Natural Language Processing API (n.d.). https://www.textrazor.com/ (accessed 20.04.2015)

Wieneke, L., Düring, M., Silaume, G., Lallemand, C., Croce, V., Lazzarro, M., Nucci, F., Pasini, C., Fraternali, P., Tagliasacchi, M., Melenhorst, M., Novak, J., Micheel, I., Harloff, E., Garcia Moron, J. (2014). histoGraph – A Visualization Tool for Collaborative Analysis of Historical Social Networks from Multimedia Collections. In L. Hughes (Ed.), *Proceedings of 18th International Conference Information Visualisation (IV), 2014 Conference*. Paris, France.

Schrifttanz: Written Dance/ Movement Poems

Susan L. Wiesner
swiesner@umd.edu
University of Maryland, United States of America

Shannon Cuykendall
scuykendall@gmail.com
Simon Fraser University, Canada

Ethan Soutar-Rau
ethan.soutar.rau@gmail.com
Simon Fraser University, Canada

Rommie L. Stalnaker
rstalnaker81@gmail.com
Bite Dance, San Diego, United States of America

Thecla Schiphorst
thecla@sfu.ca
Simon Fraser University, Canada

Karen Bradley
kbradley@umd.edu
University of Maryland, United States of America

One of the most frustrating challenges facing practitioners of Dance is the need to use spoken/written language to reference non-verbal movement. The non-verbal to verbal, and vice versa, is not a challenge isolated to practitioners of dance, but is a frustration shared amongst researchers working to represent movement through technological means. Perhaps it is enough to allow the movement, non-verbal as it is, to speak for itself. Then again, it is often through verbal language that we can make meaning from movement. The intersection of language and movement is the point at which meaning-making enforces the mind-body connection, and it is often how embodied experience is transmitted. Two HCI research projects that have studied this connection between words and movement as a means to the classification and automatic recognition of movement, ARTeFACT and POEME, are now collaborating in a new project: Schrifttanz.

ARTeFACT seeks to enable automatic recognition, identification, annotation, and retrieval of movement-based data from streaming video. A lofty goal that has been approached through the generation and analysis of both codified and abstract movement using motion capture data collected at 120 Hz with a Vicon 8-camera motion capture system and a modified Plug-In Gait full body marker set with 38 infrared reflecting markers placed on performers. Early research included instances of iterative choreography created from accelerometer data and descriptions of movement through verbal language by observers (Coartney and Wiesner, 2009). The second phase of the project involved the capture of over 200 codified dance 'steps' in various genres, the development of an ontology, and the creation of IDMove, a tool through which we were able to automatically identify and name dance movement from single dancers (Wiesner et al., 2011, 2014). The third phase ventured into the realm of identifying abstract movement that represents the conceptual metaphor CONFLICT, as introduced by Lakoff and Johnson (2003).

This data, derived from 7 dances about conflict and 19 CONFLICT terms, 396 different sections of movement, each lasting from 2-120 seconds, were captured and categorised. Critical reviews (written) about the dance works were also collected. Through statistical analysis of written and danced texts about CONFLICT, we distinguished body parts frequently used, structural preferences (e.g. stage spacing), and choreographic time spent on the different CONFLICT terms in total and per dance (Wiesner et al., 2015). The findings were validated by a small 'crowd-sourced' experiment, and movement 'rules' were developed per term in order to enable identification of a concept through movement (terms include victim, attack, surrender, struggle, conquer, hero, victory, survive, etc.) (Wiesner and Stalnaker, 2015). Concordances and collocation studies aided in the investigation of the intersections of words and movement, through the phenomenological approach of dancers' descriptions and perceptions of viewers in the form of the reviews written by dance critics. A final step has been to align the non-verbal and verbal output with concepts used in Laban Movement Analysis (Body, Effort, Shape, and Space). ARTeFACT has collected a wealth of data from captured movement, from written descriptions and articles based on the dances, and from a broad LMA perspective. In the future, various modes of comparison (e.g. the metaphor PEACE, general language studies, etc.) will further test the system.

Other researchers have incorporated LMA into their study as a means to identify movement (including studies on rats' play) (Foroud and Pellis, 2003), yet their focus is on deconstruction. POEME and ARTeFACT coincide in that rather than provide a deconstruction of the movement they both seek to characterise the essence of that movement, an approach inspired by LMA. In the case of ARTeFACT, this is accomplished by describing the relations between parts of a movement that accompany the portrayal of abstract concepts in dance. For POEME, it means finding computational means to summarize a movement experience so that it can be studied as a whole.

Where ARTeFACT has taken a more actively analytical approach - which resulted in a set of rules that describe stereotyped motion consistent with conceptual metaphors - POEME has taken a more abstract approach, letting data collection from a large number of human responses gradually expose relationships between verse and movement. POEME (Portrayal of Ephemeral Movement Experiences) is a mobile website (http://poeme.iat.sfu.ca) that interprets movement data, captured through a MARG sensor array commonly found in smartphones, into insightful, witty and whimsical poetry. The goals of POEME are twofold: 1) to create a playful game that can inspire movement exploration and 2) to explore new methods of classifying the nuances of bodily experiences through poetry and in turn understand how bodily knowledge can more easily be transmitted and articulated through words.

POEME, inspired by the Japanese form of poetry known as *haika no renga* (comic, collaborative poetry), involves the social creation of poems through turn taking (Basho, 1998). Creating these linked poems requires participants to respond to previous stanzas with original verse creating a 'movement poem', an interleaved work of verses of words and phrases of movement. We believe, as Herbelot notes, that poetry can be "analysed along the usual dimensions of prosody, syntax, semantics, etc." (2015). In POEME, each verse follows a rigid template that follows the form of a Parts of Speech Poem (PoSP). This PoSP template allows for very simple production of verses. Each poem begins with a noun, followed by two adjectives, two verbs, and an adverb. POEME composes verses from a large dictionary of English language words, based on measurements taken from prior movement responses. In an initial study, the system was trained by collecting movement responses to human- and machine-composed verses. In a later study, training was focused on two wordlists that depict different themes of movement: the 'stillness wordlist' (75 words that relate to little or no dynamic change in movement, e.g. frozen, still, serene); and a 'motion wordlist' (250 words related to the mechanics or physics of motion, e.g. buoyant, centripetal, accelerated). Using a Naive Bayes algorithm, POEME can differentiate between stillness and motion with 97% accuracy (Cuykendall et al., 2016).

Schrifttanz is a unique project that combines the scale of data collection, which POEME offers, with the nuanced language model of ARTeFACT. In Schrifttanz we explore if POEME can recognise conceptual differences in movement, specifically concepts related to CONFLICT, which have been studied in detail during the creation of the ARTeFACT system. Three dancers train the POEME agent by recording movement based on the rules defined in ARTeFACT. These recorded movement sessions are used to model conceptual relations to movement in POEME. This allows us to generalise previous findings from ARTeFACT into a new sensor modality. Also, a Laban/Bartenieff Movement Analysis using the rules-based movements establishes a secondary set of language elements and validates distinctions between movement rules. A computational agent in POEME learns to classify these movements according to the metaphor terms and their associated movements generated by the rules. We then compare POEME's ability to classify these movement metaphors with existing capabilities in ARTeFACT to validate and generalise previous findings and establish the reliability of the POEME system.

Concurrently, POEME composes verses comprised of natural language words gathered through the written articles and reviews based on dances representing the conceptual metaphors identified in ARTeFACT. Free form responses to these verses are collected using POEME's normal training procedure. A machine learning-process associates features of movement with the words. This allows POEME to generate relevant responses to movement. We explore methods of supplementing POEME's verse generation by including the conceptual metaphor model.

Words, like movements, have layers of meaning and can be recombined to create a multitude of sentences or sequences that all vary in their underlying meaning. As philosopher Mark Johnson states, "It is true that when we read, we read words. But words have meanings, and meanings go far beyond words" (2008). In Schrifttanz, the same could be said for bodily movement. POEME's use of current mobile technology and the powerful interactive engine the POEME project has developed enables us to draw from a broader population in order to study and validate the findings of ARTeFACT. Synergistically, ARTeFACT offers to POEME the ability to advance its training mechanisms while it supports the representational nature of the data generated. The two projects are pieces of a puzzle that seeks to establish a better understanding of the relationship between movement and language through the particular and the universal.

Bibliography

Basho, M. (1998). *Basho's Narrow Road: Spring and Autumn Passages.* Southbridge, MA: Stone Bridge Press, Inc.

Coartney, J. and Wiesner, S. (2009). Performance as Digital Text: capturing signals and secret messages in the media rich experience. *Literary and Linguistic Computing,* 24(2).

Cuykendall, S., Soutar-Rau, E. and Schiphorst, T. (February, 2016). POEME: A Poetry Engine Powered by Your Movement. *Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 635-64. ACM.

Foroud, A. and Pellis, S.M. (2003). The Development of "Roughness" in the Play Fighting of Rats: A Laban Movement Analysis Perspective. *Dev Psychobiol,* 42.

Herbelot, A. (2015). The semantics of poetry: A distributional reading. *Digital Scholarship in the Humanities,* 30(4).

Johnson, M. (2008). *The meaning of the body: Aesthetics of human understanding.* Chicago: University of Chicago Press.

Lakoff, G. and Johnson, M. (2003 (2nd ed.)). *Metaphors We Live By*. Chicago: University Of Chicago Press.

Simpson, T., Wiesner, S., and Bennett, B. (2014). Dance Recognition System Using Lower Body Movement. *Journal of Applied Biomechanics,* 30(1).

Wiesner, S., Bennett, B., and Stalnaker, R. (2011). ARTeFACT Movement Thesaurus. White Paper, NEH Office of Digital Humanities.

Wiesner, S. and Stalnaker, R. (2015). Representing Conflict through Dance: using quantitative methods to study choreographic time, stage space, and the body in motion. Dwyer, S., R. Franks and R. Green (Eds) *With(out) Trace: inter-disciplinary investigations into time, space and the body,* Oxford: Inter-Disciplinary Press (in press).

Wiesner, S., Stalnaker, R. and Austin, A. (2015). Training the Machine: Movement, Mo-cap, and Metaphor. *Visual Aspects of Performance,* Oxford: Inter-Disciplinary Press (in press).

Flexible Community-driven Metadata with the Component Metadata Infrastructure

Menzo Windhouwer
menzo.windhouwer@meertens.knaw.nl
Meertens Institute, Netherlands, The

Twan Goosen
twan@clarin.eu
CLARIN ERIC

Jozef Misutka
misutka@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Charles University in Prague

Dieter Van Uytvanck
dieter@clarin.eu
CLARIN ERIC

Daan Broeder
daan.broeder@meertens.knaw.nl
Meertens Institute, Netherlands, The

Introduction

Many researchers, from the humanities and other domains, have a strong need to study resources in close detail. Nowadays more and more of these resources are available online. To make these resources discoverable, they are described with metadata. These metadata records are collected and made available via central catalogues. Often, resource providers want to include specific properties of the resource in the metadata. The purpose of catalogues will be more generic and addresses a broader target audience. It is hard to strike the balance between these two ends of the spectrum with one metadata schema, and mismatches can negatively impact the quality of metadata provided. The European CLARIN infrastructure (CLARIN ERIC, 2016b) was confronted with this specific problem, and designed a solution based on a flexible mechanism to build resource specific metadata schemas, potentially using domain, community or project specific terminology, out of shared components and semantics. This paper introduces this approach and the infrastructure built for it, which is applicable to any domain with the same needs.

Component Metadata

In the Component Metadata (CMD) Infrastructure (CMDI) (CLARIN ERIC, 2016c; Broeder et al., 2012) the metadata lifecycle starts with the need of a metadata modeler to create a dedicated metadata profile for a specific type of resources, e.g., speech recordings (e.g., HZSK, 2015) or historical letters (e.g., Roorda, 2013). The modeler can browse and search a registry for components and profiles that are suitable or come close to meeting the requirements at hand. A component groups together metadata elements that belong together and can be potentially reused as a group in a different context, e.g., a location or a language description. Components can also group other components, e.g., the actor component can contain the general location component. The CMDI Component Registry (CLARIN, 2016a) already contains many of these general components. And these can be reused as they are or be adapted, i.e., add or remove some metadata elements and/or components. Also completely new components can be created to model the unique properties of the resources under consideration. All the needed components are combined into one profile that is specific to the type of resources, e.g., a profile for a speech recording (see Figure 1). Components, elements and values in this profile are linked to a semantic description, e.g., a concept, to make their meaning explicit. This feature allows the use of community specific terminology, while still creating a shared semantic layer that can be exploited by generic tools. Finally, metadata creators can create records for specific resources that comply with the profile relevant for the resource type, and these records can be provided to local and global catalogues. Notice also that CMD leaves the final responsibility of how heavy a metadata description should be to the modeler: it is perfectly possible to create a lean profile, resulting in lightweight records and combine it with a full-text index of the resources for discovery.

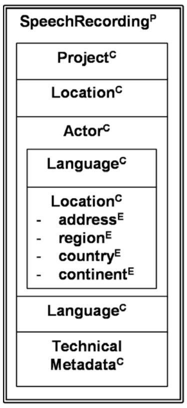


Figure 1: Example CMD profile (p), components (c) and elements (e)

The Component Metadata approach is currently being standardized by ISO Technical Committee 37. And the first part (ISO 24622-1, 2015) of this family of standards has been released and specifies the model just described.

Component Metadata Infrastructure

CLARIN built an infrastructure (see Figure 2) around the approach described in the previous section. This infrastructure is open source and provides many tools, which can readily be reused by other communities.

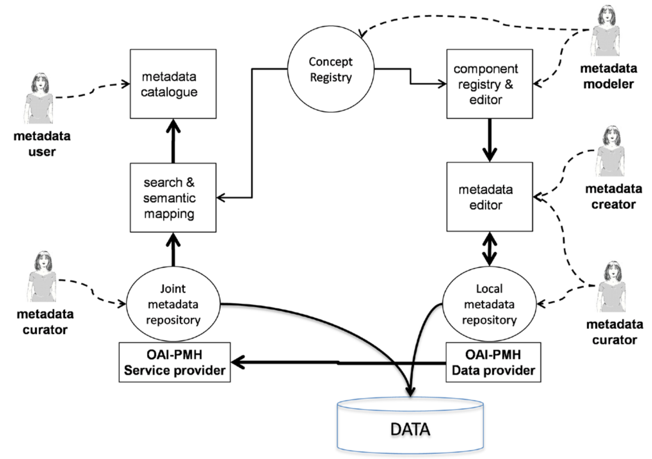


Figure 2: An overview of tools and roles in CMDI

CMDI toolkit (CLARIN, 2016b): A set of XML schemas and transformations that implement the workflow from validating component and profile specifications to conversion into profile specific XML schemas used to validate specific CMD records.

Component Registry (CLARIN, 2016b): A registry storing the profiles and components created by the community for reuse. It also provides an editor for the creation of new profiles and components and derivatives of existing ones. The backend also provides REST services based on the toolkit, i.e., serves the XML schema representation of a CMD profile.

Concept Registry (CLARIN, 2016c; Schuurman et al., 2015): A SKOS-based registry storing the communities widely accepted concepts and their relations, which form the general semantic network overlaying the specific metadata profiles (Durco and Windhouwer, 2014).

CMD editors and forms: Various general CMD editors have been developed, e.g., the desktop tool Arbil (The Language Archive, 2016a) and the online editor COMEDI (CLARINO, 2016). Also dedicated forms for specific profiles, e.g., the CMDI maker (CLARIN-D, 2016), which can be inspiring.

Repository systems: Many CLARIN centers have deployed and configured generic repository systems to store their resources accompanied by CMD records. LINDAT (UFAL, 2016) and The Language Archive (2016b) have done so as well and released their solutions as general CMD-capable repository systems.

OAI harvester (The Language Archive, 2016c): CMD records can be harvested with any OAI harvester, but this harvester is special in that has easy provisions to add transformations to CMD that enable the harvesting of other metadata formats.

Catalogues: Faceted browsing is a suitable and commonly applied method for exloring vast collections based on some key properties. The CLARIN Virtual Language Observatory (VLO) (CLARIN, 2016d) is such a browser. Although the front-end is rather CLARIN specific, the VLO importer in the back-end, which takes the harvested CMD records, determines the facets and their values and stores these in a SOLR index, is generic because the facet mapping is highly configurable and exploits the shared semantic layer. The Meertens Institute has developed an alternative faceted browser (Meertens Institute, 2016) for CLARIN-NL. The importer of this browser does not take any configuration and dynamically creates facets for any semantically different context it finds in the CMD records.

Convertors: CMDI is currently XML oriented, but other representation formats are possible. The CMD2RDF service (CLARIN-NL, 2016) provides a RDF representation to link CMD records with the world of Linked (Open) Data. Also convertors for other metadata formats, e.g., MODS or OLAC, to CMD are available.

Although this list is far from complete it shows that the CMD Infrastructure is a thriving and versatile software ecosystem. Also many parts of it are configurable, which makes it adaptable to other domains.

Lessons learned by CLARIN

When CLARIN started development on the CMD Infrastructure in its preparatory phase many things were not clear yet and a lot of flexibility was needed. Common and reusable components and concepts still had to be specified. This has lead to situations where sometimes several alternatives have coexisted for a long time in the CMD ecosystem, which can make it hard for users, especially novices, to select which one to use. It is better to prevent this kind of confusion. CLARIN advises new communities that start using the CMD Infrastructure to first setup a set of basic recommended or even obligatory components and concepts, so a stable generic core is available to the community to extend with their specific needs.

Future work

CLARIN keeps on working actively on the CMD Infrastructure, with the next version, i.e., CMDI 1.2 (Goosen et al., 2015) currently under development. The following topics will be addressed by this update:

- Lifecycle management of components and profiles.
- Connection to vocabulary services.
- Annotation of profiles and components with hints targeted at tools.

Also improvement of the metadata quality is an ongoing process. The ongoing CLARIN-PLUS project (CLARIN ERIC, 2016a) includes the design and development of tools and and a workflow that can be used used by metadata curators for quality assessment and curation of CMD records (King et al., 2015).

Conclusions

CLARIN has built a highly flexible and versatile infrastructure for metadata that is able to meet both the generic needs of catalogues and the specific needs of resource providers. The fruits of these efforts are ready to be picked by any community experiencing the same needs. CLARIN is happy to share their experiences as well as the sometimes hard learned lessons.

Bibliography

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T. and Trippel, T. (2012). CMDI: a Component Metadata Infrastructure. *Proceedings of the Metadata 2012 Workshop on Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources*. Istanbul, Turkey: European Language Resources Association (ELRA).

CLARIN (2016b). *CMDI Toolkit*http://infra.clarin.eu/cmd/ (accessed 24 February 2016).

CLARIN (2016a). *Component Registry*https://catalog.clarin.eu/ds/ComponentRegistry (accessed 24 February 2016).

CLARIN (2016c). *Concept Registry*http://www.clarin.eu/conceptregistry (accessed 24 February 2016).

CLARIN (2016d). *Virtual Language Observatory*https://vlo.clarin.eu/ (accessed 24 February 2016).

CLARIN-D (2016). *CMDI Maker*http://cmdi-maker.uni-koeln.de/ (accessed 24 February 2016).

CLARIN ERIC (2016a). *Factsheet: CLARIN-PLUS*https://www.clarin.eu/node/4213 (accessed 3 March 2016).

CLARIN ERIC (2016b). *CLARIN Infrastructure*http://clarin.eu/ (accessed 24 February 2016).

CLARIN ERIC (2016c). *Component Metadata*http://www.clarin.eu/content/component-metadata (accessed 24 February 2016).

CLARIN-NL (2016). *CMD2RDF*http://catalog.clarin.eu/ds/cmd2rdf (accessed 24 February 2016).

CLARINO (2016). *COMEDI:: The COmponent Metadata EDItor*http://clarino.uib.no/comedi/page (accessed 24 February 2016).

Durco, M. and Windhouwer, M. (2014). The CMD Cloud. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Durco, M. and Schonefeld, O. (2015). CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. *Selected Papers from the CLARIN 2014 Conference*. (Linköping Electronic Conference Proceedings). Soesterberg, The Netherlands: Linköping University Electronic Press, Linköpings universitet.

HZSK (2015). *SpokenCorpusWithResourcesProfile*https://catalog.clarin.eu/ds/ComponentRegistry/ - /?itemId=clarin.eu%3Acr1%3Ap_1442920133048&registrySpace=public (accessed 3 March 2016).

ISO 24622-1 (2015). *Language Resource Management - Component Metadata Infrastructure (CMDI) - Part 1: The Component Metadata Model*. International Organization for Standardization.

King, M., Ostojic, D. and Durco, M. (2015). Variability of the Facet Values in the VLO - a Case for Metadata Curation. *CLARIN Annual Conference 2015 - Book of Abstracts*. Wroclaw, Poland.

Meertens Institute (2016). *Search Resources and Tools at the Meertens Institute*http://www.meertens.knaw.nl/cmdi/search (accessed 24 February 2016).

Roorda, D. (2013). *CorrespondenceHistorical*https://catalog.clarin.eu/ds/ComponentRegistry/ - /?itemId=clarin.eu%3Acr1%3Ap_1360230992133&registrySpace=public (accessed 25 February 2016).

Schuurman, I., Windhouwer, M., Ohren, O. and Zeman, D. (2015). CLARIN Concept Registry: the new semantic registry. *CLARIN Annual Conference 2015 - Book of Abstracts*. Wroclaw, Poland.

The Language Archive (2016a). *Arbil*https://tla.mpi.nl/tools/tla-tools/arbil/ (accessed 24 February 2016).

The Language Archive (2016b). *FLAT*https://github.com/TheLanguageArchive/FLAT (accessed 24 February 2016).

The Language Archive (2016c). *A Simple Java Application for Managing an OAI-PMH Harvesting Workflow*https://github.com/TheLanguageArchive/oai-harvest-manager (accessed 24 February 2016).

UFAL (2016). *LINDAT/CLARIN Digital Repository Based on DSpace*https://github.com/ufal/lindat-dspace (accessed 24 February 2016).

The Manuscripts of David Livingstone and New Frontiers for Spectral Imaging

Adrian S. Wisnicki
awisnicki2@unl.edu
University of Nebraska-Lincoln, United States of America

Ashanka Kumari
ashanka.kumari@louisville.edu
University of Louisville, United States of America

Introduction

This presentation will focus on the results of the newest phase of the NEH-funded Livingstone Spectral Imaging Project (2010-present). This project seeks to apply spectral imaging and processing techniques to the study of some of the most damaged manuscripts produced by David Livingstone (1813-73), the famous Victorian traveler, abolitionist, geographer, and missionary. Our project asks whether spectral imaging can indeed restore erased or otherwise invisible portions of Livingstone's writing. More recently, we also explore whether the technology can illuminate the specific circumstances of the production and preservation of Livingstone's manuscripts, thereby better revealing the links between these manuscripts and the unique historical events that shaped the material dimensions of these manuscripts.

Methodology

Spectral imaging, a technology beginning to make a significant impact in humanities research (see bibliography), relies on imaging an object, such as a manuscript, under multiple wavelengths of light, ranging from ultraviolet (UV) through the visible color spectrum to the near-infrared. A high-resolution monochromatic digital camera automatically photographs each illumination. Imaging scientists then manipulate this raw image data with computers by applying various processing algorithms with the goal of enhancing features of interest. Often these features are made more visible by creating pseudocolor (false-color) representations of the object that foreground or suppress other object elements.

Our previous research (Wisnicki 2011) applied spectral imaging to recover the text of a diary that Livingstone had written over newspaper pages and that had become illegible because, first, the considerable fading of Livingstone's ink (which he had concocted out of a local African clothing dye) and, second, the continuing prominence of the black newsprint over which Livingstone wrote his text. Using a combination of new and established spectral image processing techniques, our team successfully suppressed the newsprint so that it no longer interfered with Livingstone's text. By enhancing his writing so that it could be read easily, we were able to recover some 99% of Livingstone's words (up from about 40% before) and, in turn, use this technique to reveal important new information about Livingstone's strategies for representing his experiences in interacting with local populations in Central Africa.

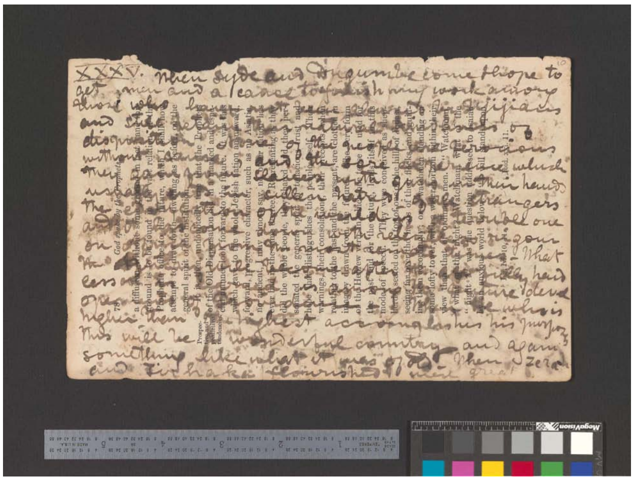


Figure 1. Page XXXV of Livingstone's 1870 Field Diary in natural light. Copyright National Library of Scotland. CC BY-NC 3.0

Our current project takes our previous research in a new direction to suggest that spectral imaging can illuminate the production and preservation history of Livingstone's fragmentary 1870 Field Diary. With spectral imaging, we have enhanced or revealed material features of this diary, such as overwriting, staining, and page topography, that are often not visible or difficult to discern with the naked eye. Our results represent an important advance in the study of manuscripts with spectral imaging, particularly because previous related research has generally focused on using of spectral imaging for the recovery of faded texts or illegible palimpsests such as the Dead Sea Scrolls (Shor 2012) and the Archimedes Palimpsest (Netz and Noel 2007).

Historical and Scholarly Implications

Livingstone wrote the 1870 Field Diary in central Africa and carried it with him to his death in 1873 in what is now Zambia. Livingstone's supporters then transported his manuscript (and his corpse) from the interior to the coast of Africa and back to Britain, where different repositories have since preserved the manuscript. The pages of the diary, therefore, bear the remnants of this complex history, from the changes and deletions that Livingstone made over time, to the marks left on the manuscript pages by the diverse local African environments through which the diary traveled, to the traces of modern curatorial methods used to preserve the diary's pages.

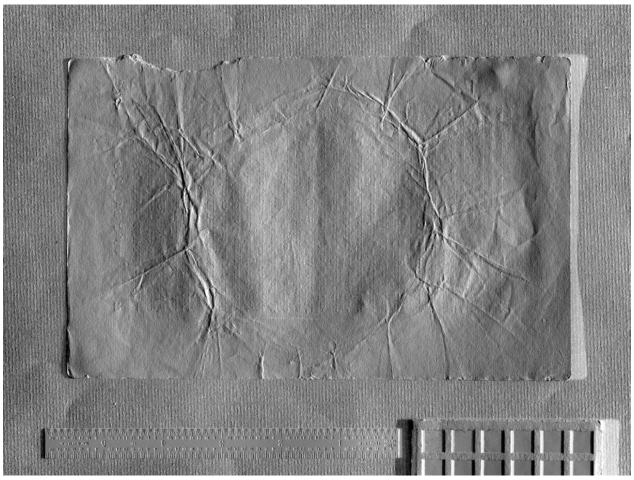


Figure 2. Page XXXV of Livingstone's 1870 Field Diary with topography illuminated by spectral image processing. Copyright National Library of Scotland. CC BY-NC 3.0

Our efforts to study the material history of Livingstone's diary with spectral imaging represent an important intervention in the field. Our work demonstrates new methods and frontiers for the use of spectral imaging in the humanities because it underscores that spectral imaging can do more than recover faded or invisible text. Spectral Imaging can provide crucial insights into the passage of a manuscript through time and into the relationship between a manuscript and the specific historical circumstances from which it emerged. By applying this technology to the 1870 Field Diary, we have gained key insights into the strategies by which Livingstone shaped his experiences and identity for public consumption and the impact of specific moments of handling in determining material features of the manuscript. These insights promise to enhance our understanding of Livingstone's biography and the history of the many African cultures in which he worked.

Moreover, the spectral image processing techniques we have developed for the 1870 Field Diary can now be applied to study other manuscripts with notable textual and material features or, indeed, other cultural objects such as paintings and sculptures whose surfaces retain their material history and bear the marks of the many people and events that have come to shape the objects we encounter today. Our project is helping us distinguish such marks, but, in our case, the project also sets the stage for future research in which we might use the spectral signatures of specific marks as a portal to defining tangible aspects of nineteenth-century African environments from which Livingstone's manuscripts emerged and through which these manuscript circulated.

Bibliography

Chabries D. M., Booras S. W. and Bearman G. H. (2003). Imaging the past: recent applications of multispectral imaging technology to deciphering manuscripts, *Antiquity*, **77**(296): 359–72.

Easton R. L. Jr., Knox K. T., Christens-Barry W. A., Boydston K., Toth M. B., Emery D. and Noel W. (2010). Standardized system for multispectral imaging of palimpsests, *Proceedings of SPIE* 7531, *Computer Vision and Image Analysis of Art.* 75310D.111, San Jose, California.

Goltz D. M., Cloutis E, Norman L. and Attas M. (2007). Enhancement of faint text using visible (420-720 nm) multispectral imaging, *Restaurator*, pp. 11–28.

Joo Kim S., Deng F. and Brown M. S. (2011). Visual enhancement of old documents with hyperspectral imaging, *Pattern Recognition*, **44**(7): 1461–69.

MacDonald L. W., Giacometti A., Campagnolo A., Robson S., Weyrich T., Terras M. and Gibson A. (2013). Multispectral imaging of degraded parchment, *Proceedings of Computational Color Imaging, 4th International Workshop*, CCIW 2013, Chiba, Japan, 3-5 March 2013.

Marengo E., Manfredi M., Zerbinati O., Robotti E., Mazzucco E., Gosetti F., Bearman G., France F. and Shor P. (2011). Development of a technique based on multi-spectral imaging for monitoring the conservation of cultural heritage objects, *Analytica Chimica Acta*, **706**(2): 229–37.

Netz R. and Noel W. (2007). *The Archimedes Codex: How A Medieval Prayer Book Is Revealing The True Genius of Antiquity's Greatest Scientist.* London: Da Capo Press.

Shor, P. (2012). *The Leon Levy Dead Sea Scrolls Digital Library*. Israel Antiquities Authority. Retrieved from http://www.deadseascrolls.org.il/home.

Walters Art Museum. (2008). *The Archimedes Palimpsest data set*. Baltimore: Walters Art Museum. Retrieved from http://www.archimedespalimpsest.org/.

Wisnicki, A. S. (2011). *Livingstone's 1871 Field Diary: a multispectral critical edition*. Beta edition; first edition; corrections. Los Angeles: UCLA Digital Library. Retrieved from http://livingstone.library.ucla.edu/1871diary/index.htm.

Public and Private Views of Texts in Digital Editions – The Case of the Kanseki Repository

Christian Wittern
cwittern@gmail.com
Kyoto University, Japan

Introduction

The established pattern for a scholarly digital edition today is the website, which in many cases has a unique and well thought out user interface. It concentrates all information pertaining to the topic, but allows little interaction of the reader with the texts beyond what has been designed into the user interface by the developers of the site.

Although there are also many efforts to go beyond this and experiment with new forms of reading in the digital age, for example protocols for sharing annotations or reading communities, they have not yet reached a stage where they would be available to mainstream researchers.

In contrast to the polished and well advertised flagship editions of digital projects at prestigious institutions, there is also a continuing trend of making the source texts that feed into digital editions available in a way that not only allows, but actively encourages tinkering with the texts. The distribution for this latter type of texts is frequently on the site github.com, which is the world largest repository for software development[1], but as a free site for data sharing with collaborative editing it is becoming more popular for other purposes as well.

Some of the large scale text projects using Github in this way include the *Chinese Buddhist Electronic Text Association* (CBETA)[2], *gitenberg*[3] and *EEBO-TCP*[4] to name just a few.

The presentation proposed here reports on the *Kanseki Repository*, a project that tries to establish a link between these two different types of text dissemination and through this combination to achieve the best of the two worlds: The texts can still be presented through a sophisticated web interface, which uses Github as a backend storage and reads the texts from there for presentation to the user. These texts can be forked (that is, cloned and copied to the user's account) which makes it possible for the user to revise or annotate them. The system is set up in a way that it will show this private copy of the text if configured so by the user.

Methodological considerations

The goal here is not simply to upload as many texts at possible to public archives, but rather to develop a platform that allows the basic tenets of scholarly editing to co-exist and thrieve with the possibilities and affordances of the digital medium. It is therefore of paramount importance, to consider the requirements as voiced from practioners of scholarly editing and implement them as transparently as possible.

Record and interpretation

In a seminal article, the Swiss scholar Hans Zeller (1995) emphasised the fact that all scholarly editing should make a clear distinction between the record of what is transmitted and the scholarly interpretation thereof. While this distinction is blurry at times, it has informed the design of the *Kanseki Repository*, which arranges the editions of a text it represents into those that strive to faithfully reproduce a text according to some textual witness ('record') and those that critically consider the content and make alteration to the text by adding punctuation, normalizing characters, collating from other evidence etc. ('interpretation').

Additional requirements

Peter Shillingsburg (2015) outlines the following requirements of a digital edition (slightly edited for clarity):

a. Digitize images of all the documents. That will make it possible, from anywhere in the world, to see any document side-by-side with any other document without traveling from Tokyo to Marburg and New York.

b. Prepare a table of variants to show how all the documentary texts differ from one another.

c. Write a textual history that explains the relationships among the variant documents and explains why we should care–why it is important to know.

d. Transcribe at least one of the documents so that the variants list can be more easily used. Or transcribe all the documents so that readers can select and read any one. Transcribing all the documents will also make machine collation possible.

e. Edit one of the transcriptions to correct obvious errors. This will preserve the text as a historical documentary text but will help readers avoid the distractions caused by scribal or compositorial errors.

The architecture developed in this project strives to implement as much of this as seems feasible within the limits of the current funding and other constraints. The components will be further described below.

Main components of the project architecture

The architecture of the project consists conceptually of two parts: (1) the text repository and (2) clients accessing this repository. Of these clients there are currently two, both developed within the project.

Backend: Repository of texts: github.com/kanripo

Since every text has its own unique history and witnesses, every text is allocated to its own repository (in the technical sense). These repositories, currently more than 9000, are are organized according to the traditional Chinese cataloging principles and kept under the Github account @kanripo. Since the texts are publicly accessible, they can be freely downloaded and cloned even without ever touching the client interfaces. Due to the large number of texts and the Github interface, which seems foreign and intimidating to most readers of classical Chinese texts, special clients cover most needs in interfacing with the texts.

Client interfaces

Web interface at www.kanripo.org

This website provides access to the texts, including full-text search, display of transcribed text and facsimile(s) of different editions. Users can log in using their Github credentials and get access to more advanced functions such as selecting lists of text of special interest, advanced sorting functions by text category or date as well as cloning of texts to the Github user account and editing on site. The site went into testing mode in October 2015 and a first public release has taken place in March 2016.

Mandoku, a stand alone tool for further immersive reading and study

In addition to the website, an Emacs module called Mandoku (see Wittern, 2012, 2013, 2014a, 2014b) has been developed (as an extension to Orgmode, see Domnik et al., 2015), that uses the API of the website to provide the same search functions, but clones the texts of interest to the user to her local machine, thus providing advanced editing possibilities and offline access.

Towards a platform for text-based Chinese studies

All modes of interaction described above are based on the distributed version control system git, using the Github site as a 'cloud storage'. However, in addition to providing storage, Github also provides a feedback mechanism through "pull-requests", where users can flag corrections to a text for the @kanripo editors to consider for inclusion in the canonical version, thus making it available to all users.

The model outlined here is extensible and allows other developers of websites related to Chinese studies to access the same texts, and provide specialized services to the user, for example by enhancing the text through NLP processing. These enhanced versions can be saved ("committed" in git language) in the same way to the users account and are then also visible to the client programs described here.

This will open the door to a open platform of texts for Chinese studies, where the texts of interest to the users form the center of a digital archive, with different services and analytical tools interacting and enhancing it. The user, who makes a considerable investment in time and effort when close reading, researching, translating and annotating the text, never loses control of the text and does not need to worry about losing access to it when one of the websites goes offline.

By providing versioned access to the texts in question, it is also possible to make any analytical results reported in research publications reproducible (Rawal, 2015) by indicating the additional tools and processes needed, ideally also in a Github repository in the same ecosystem.

The aim is not just to provide a static, completed, definitive edition of a text, but as fertile a ground as possible for the interaction between the text and its readers, hopefully improving both through this process.

Bibliography

Domnik, C. et al. (2015). Org mode for Emacs — Your Life in Plain Text, at http://orgmode.org (accessed 2016-03-02).

Rawal, V. (2015). Reproducible Research Papers using Org-mode and R: A Guide, at https://github.com/vikasrawal/orgpaper (accessed 2016-03-02).

Shillingsburg, P. (2015). Global Textual Scholarship: An American View, paper delivered at a symposium on textual studies in Japan available online at http://sunrisetc.blogspot.jp/2015_04_01_archive.html (accessed 2016-03-02).

Wittern, C. (2012). Text Representation and Interchange in the Digital Age, paper delivered at Annual conference of the Japanese Association for Digital Humanities 2012 at University of Tokyo, Sep. 15-17, 2012.

Wittern, C. (2013). Beyond TEI: Returning the Text to the Reader. *Journal of the Text Encoding Initiative* [Online], 2013, 4. (http://jtei.revues.org/691) (accessed 2016-03-02).

Wittern, C. (2014a). Kanripo and Mandoku: Tools for git-based distributed repositories for premodern Chinese texts. In *Digital Humanities 2014 Book of Abstracts*, pp. 408-409.

Wittern, C. (2014b). Conventions for a repository of premodern Chinese texts. In: 東洋学へのコンピュータ利用第２５回セミナー，２０１４年３月１５日, p. 73-88.

Zeller, H. (1995). Befund und Deutung - Interpretation und Dokumentation als Ziel und Methode der Edition. In: Martens, G. and Zeller, H. (ed.), *Texte und Varianten : Probleme ihrer Edition und Interpretation*. München, pp. 45-89, translated as Record and Interpretation: Analysis and Documentation as Goal and Method of Editing. In: Gabler, H., Bornstein, G. and Pierce, G. B. (ed.), *Contemporary German Editorial Theory*, Ann Arbor, pp. 17-58.

Notes

[1] The site claims: "GitHub is where people build software. More than 11 million people use GitHub to discover, fork, and contribute to over 28 million projects." (http://www.github.com, accessed 2016-03-02).

[2] https://github.com/cbeta-org/xml-p5 (accessed 2016-03-02), more than 4000 Chinese Buddhist texts in TEI P5 markup.

[3] https://github.com/GITenberg (accessed 2016-03-02); a project that uploaded more than 40000 electronic texts from the Gutenberg.org project to make it possible to improve the texts that are otherwise distributed without the possibility for the readers to provide corrections.

[4] https://github.com/textcreationpartnership (accessed 2016-03-02), which contains 25000 texts released to the public on Jan. 1, 2015.

Oulipian Stylometry

Mark Wolff
wolff.mark.b@gmail.com
Hartwick College, United States of America

The Oulipo, or Ouvroir de littérature potentielle, is a group of writers in Paris who for over fifty years have experimented with algorithmic techniques for writing and reading literature. Raymond Queneau, one of the group's co-founders, proposed a method in 1964 for analyzing the syntactic structure of texts written in French, and he believed that the method, which he called matrix analysis, could provide a measure of an author's style.

What can matrix analysis contribute to stylometry? Apart from its origins as a form of computational play for play's sake (Wolff, 2007), matrix analysis offers an alternative to lexically based techniques for authorship attribution such as Burrow's Delta (2002). Rybicki and Eder reported that Delta does not work as well for texts written in French as for those in English and German (2011). Antonia, Craig and Elliott have shown that analyzing the frequencies of lexical n-grams where n > 1 does not usually yield very good results (2014). Researchers have developed syntactical methods based on bigrams of labels for a simplified parsing of texts (Hirst and Feiguina, 2007) and on correlations between semantics and the structures of dependent and independent clauses (Allison et al., 2013). In the latter study the researchers concluded that their definition of style "entailed [...] *a method for looking for it*" (28). The early Oulipo would have agreed with this approach. Recognizing that an author aware of how he or she used words syntactically might apply Queneau's matrix analysis to change his or her "manner," François Le Lionnais (the other co-founder) claimed that matrix analysis could serve as a "literary prosthesis" exemplifying the vocation of the group (Bens, 2005: 246). For Le Lionnais, the most important focus of the Oulipo should be the synthesis of new possibilities for understanding literary phenomena (Oulipo, 1973: 17). Matrix analysis enables the identification of significant syntactical patterns for further inquiry into an author's style. These patterns would be difficult to ascertain without a method like the one developed by Queneau.

Queneau devised a grammatical schema of the French language for describing the construction of word pairs using a system similar to linear algebra (1964). He began by dividing all elements of speech into two categories: signifiers, which include nouns, adjectives, and verbs (except *avoir* and *être*); and formatives, which include everything else (*avoir*, *être*, pronouns, articles, conjunctions, prepositions, adverbs, interjections, etc.). Given a sentence, one can construct two matrices where the first matrix contains all formatives and the second all signifiers:

Figure 1

$$\left\| \begin{matrix} le & a & la \end{matrix} \right\| \times \left\| \begin{matrix} chat \\ mangé \\ souris \end{matrix} \right\| = (Le \times chat) + (a \times mangé) + (la \times souris)$$

If a sentence contains two consecutive formatives or signifiers, one can use a unitary element to construct the matrices:

Figure 2

$$\left\| \begin{matrix} Le & 1 & a & bien & la & 1 \end{matrix} \right\| \times \left\| \begin{matrix} vilain \\ chat \\ 1 \\ mangé \\ belle \\ souris \end{matrix} \right\| = \begin{matrix} (Le \times vilain) + (1 \times chat) + (a \times 1) + \\ (bien \times mangé) + (la \times belle) + (1 \times souris) \end{matrix}$$

By adopting the conventions that neither (1 × 1) nor (Y × 1) + (1 × Z) are allowed within a sentence, one avoids uninteresting or redundant word pairs.

Queneau proposed observing the distribution of formatives and signifiers in a text using the relation F + Uf = S + Us, where F is the number of formatives, S the number of signifiers, Uf the number of unitary elements paired with signifiers (1 × Z), and Us the number of unitary elements paired with formatives (Y × 1). Noting that even if an author like Flaubert worked tirelessly to vary how he wrote, Queneau believed that this distribution could serve as a "potential" but unconscious indicator of the author's style (1965: 319).

In order to test Queneau's hypothesis, I applied his matrix analysis to a corpus of 328 nineteenth-century French novels from the ARTFL-FRANTEXT database using Helmut Schmidt's part-of-speech tagger (1995). Figure 3 is a biplot of a principal components analysis of scaled values for F, S and Uf (Us is excluded to avoid collinearity) and the graph shows that works by the authors Jules Barbey d'Aurevilly, Alexandre Dumas and Honoré de Balzac cluster separately whereas works by George Sand form distinct clusters.

Figure 3

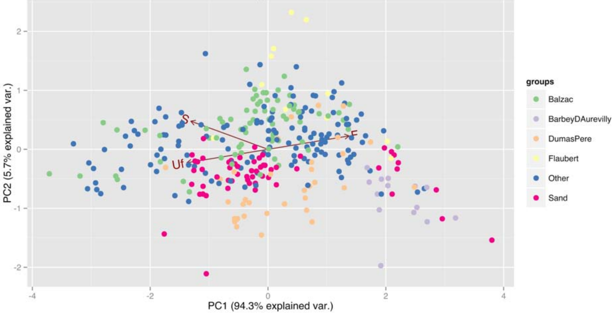
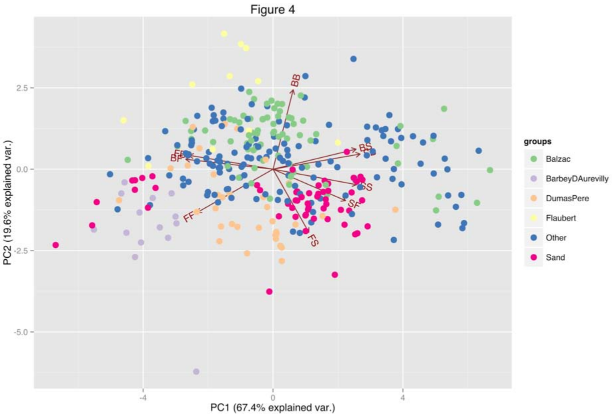


Table 1 indicates the results of using support vector machines with a radial basis function kernel to build a classification model for the texts (Kuhn, 2015). Sixty-seven percent of the corpus was used for training the model with 10-fold cross-validation. A one-against-one method was used for classifying the test set (Karatzoglou, 2004: 7-8). The results show a moderate authorial signal in the works of Barbey d'Aurevilly and weaker signals for Dumas and Balzac.

Table 1

	Sensitivity	Specificity	Prevalence	Balanced Accuracy
Class: BarbeyDAurevilly	0.750000	0.934783	0.041667	0.842391
Class: DumasPere	0.636364	0.776471	0.114583	0.706417
Class: Balzac	0.875000	0.458333	0.250000	0.666667
Class: Flaubert	0.000000	1.000000	0.031250	0.500000
Class: Sand	0.000000	1.000000	0.197917	0.500000

Figure 4



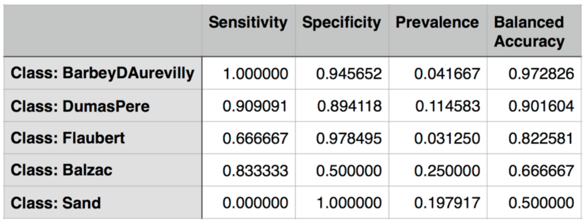
To build a better model using matrix analysis, one can observe the distribution of bigrams of matrix analysis pairs in the corpus. Given the labels **F** for (Y × 1), **S** for (1 × Z), and **B** for what Queneau called a biword (Y × Z), one can transcribe a text into a series of these letters. For instance, the sentence from George Sand's *Indiana* (1832):

> (Toute × 1) (sa × conscience), (c' × 1) (était × 1) (la × loi); (toute × 1) (sa × morale), (c' × 1) (était × 1) (son × droit).

can be represented as **FBFFBFBFFB**. With the feature set of bigrams **FF**, **FB**, **BF**, **BB**, **BS**, **SB**, **SS**, **SF**, and **FS**, one can analyze its distribution within the corpus. Figure 4 is a biplot of a principal components analysis of the corpus with the nine bigrams as variables and suggests that at least some authors do have measurable differences in style.

Table 2 shows the results of building a model with support vector machines to classify the texts by author according to the distribution of bigrams.

Table 2

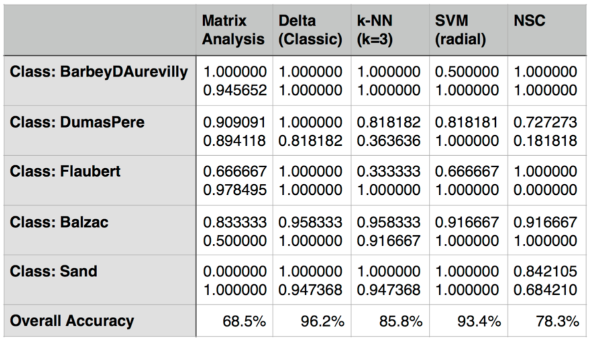


	Sensitivity	Specificity	Prevalence	Balanced Accuracy
Class: BarbeyDAurevilly	1.000000	0.945652	0.041667	0.972826
Class: DumasPere	0.909091	0.894118	0.114583	0.901604
Class: Flaubert	0.666667	0.978495	0.031250	0.822581
Class: Balzac	0.833333	0.500000	0.250000	0.666667
Class: Sand	0.000000	1.000000	0.197917	0.500000

The model can identify the styles of Barbey d'Aurevilly and Dumas with very good accuracy, and it can detect authorial signals in works by Flaubert and Balzac. The model does not do well identifying works by Sand because they seem to evince two distinct styles (as suggested by Figure 4). The small cluster of works by Sand on the left side of the graph include *François le champi*, *Elle et lui*, *Le Château des désertes*, *La Mare au diable*, *Consuelo*, *Indiana*, *Lélia*, and *La Comtesse de Rudolstadt*. The predominance of formatives in this cluster is perhaps indicative of a more conversational style: such an hypothesis would require further analysis.

Compared to other classification methods based on wordlists, matrix analysis does not deliver as high a level of accuracy. Table 3 summarizes the sensitivity and specificity of several classification models with different statistical techniques for the five selected authors using the *classify()* function of the R software package *stylo* (Eder et al., 2013):

Table 3



	Matrix Analysis	Delta (Classic)	k-NN (k=3)	SVM (radial)	NSC
Class: BarbeyDAurevilly	1.000000 0.945652	1.000000 1.000000	1.000000 1.000000	0.500000 1.000000	1.000000 1.000000
Class: DumasPere	0.909091 0.894118	1.000000 0.818182	0.818182 0.363636	0.818181 1.000000	0.727273 0.181818
Class: Flaubert	0.666667 0.978495	1.000000 1.000000	0.333333 1.000000	0.666667 1.000000	1.000000 0.000000
Class: Balzac	0.833333 0.500000	0.958333 1.000000	0.958333 0.916667	0.916667 1.000000	0.916667 1.000000
Class: Sand	0.000000 1.000000	1.000000 0.947368	1.000000 0.947368	1.000000 1.000000	0.842105 0.684210
Overall Accuracy	68.5%	96.2%	85.8%	93.4%	78.3%

To minimize the effects of semantic variation, the wordlists for classification with *stylo* were culled 100% (only those words that appear at least once in every text were included).

Despite the low accuracy of matrix analysis, it is possible to identify sample sentences that exemplify an author's style with sequences of bigrams. In Figure 4 the left group of Sand's texts clusters near the vectors for **FF**, **BF** and **FB**. Scanning the texts for the highest relative frequency of these bigrams yields sentences such as this from Sand's *Consuelo* (1842):

> (Il × faut) (que × 1) (je × sache) (comment × 1) (elle × 1) (se × tient), (ce × 1) (qu' × 1) (elle × fait) (de × 1) (sa × bouche) (et × 1) (de × 1) (ses × yeux). **BFBF FBFFBFBFFB**

The syntax of this sentence as schematized by matrix analysis contains within it the syntax of the previously quoted sentence, inviting further inquiry into how Sand constructed her texts. Although Sand most likely did not think of her writing style in the terms conceived by Queneau, matrix analysis represents a method for thinking about style that not only measures how words are used but can also inform potentially the act of writing and reading. Lexically-based techniques consider texts as "bags of words" with structure-less frequencies, but matrix analysis approaches texts as objects that have undergone a process of development. As an Oulipian procedure, matrix analysis allows the reader to detect reproduced and reproducible patterns through an interactive process of textual exploration.

Queneau's matrix analysis represents an analytical method for defining style that classifies texts according to their structure. The Oulipo in the 1960s made a distinction between the quality of works of literature and the potentiality of the methods used to create works of literature (Bens, 2005: 80). Practitioners of computational text analysis can observe a similar distinction between the accuracy of text classification and the potentiality of classification methods for understanding literature. If style implies how words are used more than what words are used, stylometry should seek to better understand how words are used. The Oulipo provides us with an example of this kind of inquiry into computationally revealed text structures. Queneau performed small experiments with matrix analysis, but Le Lionnais imagined the possibility of harnessing machines to support the necessary calculations on a larger scale (Bens, 2005: 246). Following the Oulipian notion of "plagiarism by anticipation" (Oulipo, 1973: 23), we can understand matrix analysis as a precursor of stylometry in the digital humanities.

Bibliography

Allison, S., Gemma, M., Heuser, R., Moretti, F., Tevel, A. and Yamboliev, I. (2013). Style at the Scale of the Sentence. *Literary Lab* 5. Stanford University. Retrieved from <http://litlab.stanford.edu/LiteraryLabPamphlet5.pdf>

Antonia, A., Craig, H. and Elliott, J. (2014). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. *Literary and Linguistic Computing*, 29/2: 147–63. DOI: 10.1093/llc/fqt028

Bens, J. (2005). *Genèse de l'Oulipo 1960-1963*. Bordeaux: La Castor Astral.

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17/3: 267–87. DOI: 10.1093/llc/17.3.267

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: *Digital Humanities 2013: Conference Abstracts.* University of Nebraska--Lincoln, NE: 487-89.

Hirst, G., and Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22/4: 405–17. DOI: 10.1093/llc/fqm023

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11/9: 1–20.

Kuhn, M. (2015). *caret: Classification and Regression Training*. Retrieved from <http://CRAN.R-project.org/package=caret>

Oulipo (1973). *La Littérature potentielle (Créations Ré-créations Recréations)*. Paris: Gallimard.

Queneau, R. (1964). L'Analyse matricielle du langage. *Etudes de linguistique appliquée*, 3: 37–50.

Queneau, R. (1965). *Bâtons, chiffres et lettres*. Paris: Gallimard.

Rybicki, J. and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26/3: 315–21. DOI: 10.1093/llc/fqr031

Schmid, H. (1995). TreeTagger: a language independent part-of-speech tagger. Retrieved from <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

Wolff, M. (2007). Reading Potential: The Oulipo and the Meaning of Algorithms. *Digital Humanities Quarterly*, 1/1.

DIVAServices-Spotlight – Experimenting with Document Image Analysis Methods in the Web

Marcel Würsch
marcel.wuersch@unifr.ch
University of Fribourg, Switzerland

Michael Bärtschi
michael.baertschi@unifr.ch
University of Fribourg, Switzerland

Rolf Ingold
rolf.ingold@unifr.ch
University of Fribourg, Switzerland

Marcus Liwicki
marcus.liwicki@unifr.ch
University of Fribourg, Switzerland

Introduction

We present an easy-to-use web-based user interface which allows scholars working on manuscripts to assess the usefulness of automatic document image analysis (DIA) methods when incorporating automatic processes into their workflows. In contrast to existing web interfaces (Clausner, Pletschacher, and Antonacopoulos, 2011; Embach et al., 2013), this interface allows the user to directly upload images of the manuscript of interest without any registration. Thus, a fast assessment of a variety of algorithms and DIA processes can be performed. DivaServices is not a specialized tool for one specific use-case, but a collection of tools for several tasks in different use cases.

With this web interface we build on our previous initiative (Würsch, Ingold, and Liwicki, 2015) to provide access to a wide range of DIA methods to the research communities in Computer Science and the Humanities. While the existing DivaServices are already useful to integrate state-of-the-art DIA methods into new research applications it is still difficult for researchers with little programming experience to estimate the capabilities of the offered methods. For example, it is not easy to know which binarization, text line segmentation, or OCR method and which parameters would work best on a given manuscript.

In order to overcome this shortcoming, we present a web application that allows interacting with all offered methods. Users are able to upload their own images, perform experiments on them and have the results visualized. With this, researchers in the Humanities should get a better understanding on what the methods developed in the Computer Science community are able to achieve. The other way round, researchers from the Computer Science will have their methods exposed to a much broader range of data and can gather feedback to further improve the methods. This feedback loop should enhance communication between the two communities such that future methods can target the respective needs even better.

DivaServices-Spotlight is built in a highly modular way, providing Graphical User Interface (GUI) blocks for various kinds of input and output parameters. The web interface is therefore automatically updated when new methods are added without the need of any further ado. Furthermore, based on our continuous open source support, this tool is available under a LGPL v2.1 license and the source code can be downloaded from github.[1]

DivaServices-Spotlight

DivaServices-Spotlight[2] is a web application that allows user to upload their own image data, perform experiments and investigate the results. This should help in deciding whether an algorithm can help solving a particular problem or not, and which parameters are best for the data at hand. Figure 1 provides an example of such an experiment where the highlighted area (left) is segmented into text lines (right) and visualized for the user.

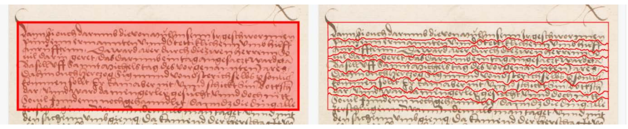


Figure 1 Example of an executed experiment using DivaServices. The highlighted region (left) is segmented into separate text lines (right) and visualized for the user.

In this part we give an overview of how the different user interface works and provide an example on how to perform a workflow.

The User Interface

From the welcome page the three main parts of DivaServices-Spotlight are available: Images, for uploading and manipulating input images; Algorithms, for executing methods; and Results, for accessing computed results.

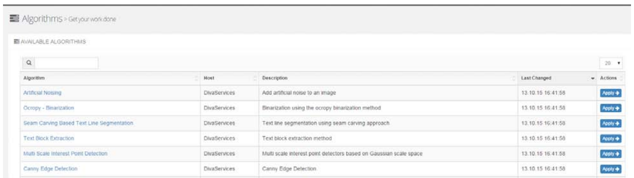
Images

Via the "Images" link the user can view his already uploaded images (Gallery) or upload new images (Upload). In its current version all uploaded images are automatically converted into the PNG format and users can upload a maximum of ten images at the same time.

Users get the possibility to apply various pre-processing steps onto their image. It is possible to crop an image to a specific size and values such as *brightness*, *contrast*, and *saturation* can be adjusted. Performing these pre-processing steps can lead to better results of varying methods.

Algorithms

On the "Algorithm" page, all currently available methods are listed (c.f. Figure 2). When selecting "Apply" on a given method, the user is asked to select one of his uploaded images.



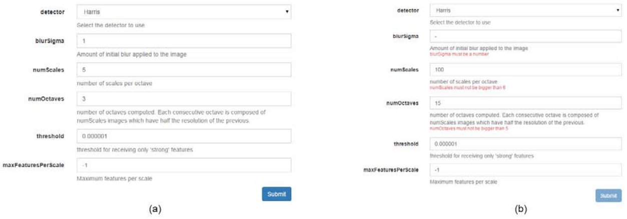
Algorithms > Get your work done

AVAILABLE ALGORITHMS

Algorithm	Host	Description	Last Changed	Actions
Artificial Noising	DivaServices	Add artificial noise to an image	13.10.15 16:41:58	Apply
Ocropy - Binarization	DivaServices	Binarization using the ocropy binarization method	13.10.15 16:41:58	Apply
Seam Carving Based Text Line Segmentation	DivaServices	Text line segmentation using seam carving approach	13.10.15 16:41:58	Apply
Text Block Extraction	DivaServices	Text block extraction method	13.10.15 16:41:58	Apply
Multi Scale Interest Point Detection	DivaServices	Multi scale interest point detectors based on Gaussian scale space	13.10.15 16:41:58	Apply
Canny Edge Detection	DivaServices	Canny Edge Detection	13.10.15 16:41:58	Apply

Figure 2 The "Algorithms" page provides an overview of all available methods with a short description of what they can be used for. Using the "Apply" button one method can be used.

On the page for a specific algorithm the user then has to specify input for this method. The input elements are created automatically based on the specifications of the method. For certain input elements a method can also specify ranges of possible values. The input of the user is validated and error messages are displayed should the input be not in a valid range.



(a)

detector: Harris
Select the detector to use

blurSigma: 1
Amount of initial blur applied to the image

numScales: 5
number of scales per octave

numOctaves: 3
number of octaves computed. Each consecutive octave is composed of numScales images which have half the resolution of the previous.

threshold: 0.000001
threshold for receiving only 'strong' features

maxFeaturesPerScale: -1
Maximum features per scale

Submit

(b)

detector: Harris
Select the detector to use

blurSigma: -
Amount of initial blur applied to the image
blurSigma must be a number

numScales: 100
number of scales per octave
numScales must not be bigger than 6

numOctaves: 15
number of octaves computed. Each consecutive octave is composed of numScales images which have half the resolution of the previous.
numOctaves must not be bigger than 5

threshold: 0.000001
threshold for receiving only 'strong' features

maxFeaturesPerScale: -1
Maximum features per scale

Submit

Figure 3 Different input types and validations. DivaServices-Spotlight offers automatic generation of input blocks for different types of inputs (a) like numbers, strings, and selection. Automatic validation (b) ensures that the user input is within ranges specified by the method.

Figure 3 (a) shows how various input blocks are generated by DivaServices-Spotlight. Currently it is possible to generate blocks for the following elements: strings (textual data), numbers, selection (one of multiple), and checkboxes. In Figure 3 (b) validation of input elements is visualized. When the user inputs data that is not valid for the given input type (e.g. text data for numbers) an error message is displayed and the user cannot execute the method.

Furthermore, an algorithm can specify that a user needs to select a region within the image. This is needed by methods which want to only work on a subset of the image and can speed up the runtime, as well as the quality of the results of a method (e.g., of text line detection). DivaServices-Spotlight allows for drawing the following selections onto an input image: rectangle, polygons, and circles. These regions are drawn using the mouse. Rectangles and circles can be created using a simple click and drag operation. Polygons are created through manually creating every point of the polygon and clicking near the start point to close it. After creating the various highlighters, they can be edited (e.g., a single point of a polygon can be moved to a new location after creation). The various highlighters are visualized in Figure 4.

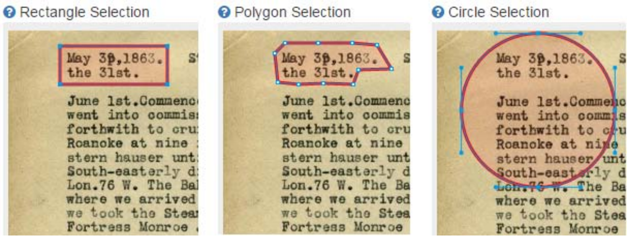
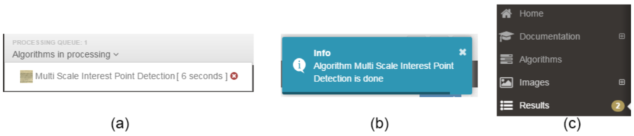


Figure 4 The different selection methods; rectangle (left), polygon (center), and circle (right).

Once the user has entered necessary parameters and selected a region on the image (if needed) the execution can be started using "Submit". The user is notified of the process at the top of the page that shows more information when clicked on with the mouse (Figure 5 (a)). Once the execution is finished, again the user is notified by a small balloon that pops up in the top right corner (Figure 5 (b)). Also, the counter behind the “Results” link in the menu navigation is increased (Figure 5 (c)).


PROCESSING QUEUE: 1
Algorithms in processing
Multi Scale Interest Point Detection [6 seconds]

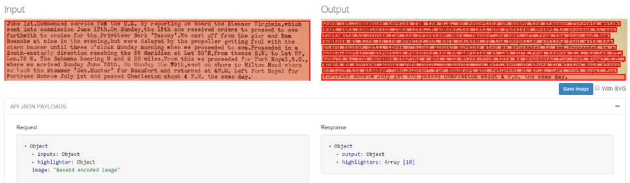
Info
Algorithm Multi Scale Interest Point Detection is done

(a) (b) (c)

Figure 5 Notifications shown to the user about the current status of an execution (a), when an algorithm finishes (b), and the number of available results (c).

Results

The “Results” page provides the user an overview of all available results. Using the “+” button on a specific result will show him the computed result. On the left side the user sees the input image as well the used parameters and on the right side the user gets a visualization of the results.


Input

June 1st.Commenced service for the U.S. by reporting on board the Steamer Virginia,which
went into commission June 13th.On Sunday,the 15th she received orders to proceed to sea
forthwith to cruise for the Privateer Bark 'Tacony'.We cast off from the pier and Ram
Roanoke at nine in the evening,but were delayed by the propeller getting foul with the
stern hauser until three o'clock Monday morning when we proceeded to sea.Proceeded in a
South-easterly direction reaching the 68 Meridian at Lat 30'N,from thence S.W. to Lat 27,
Lon.76 W. The Bahamas bearing W and S 20 miles.From this we proceeded for Port Royal,S.C.,
where we arrived Sunday June 28th. On Monday the 29th,went on shore to Hilton Head where
we took the Steamer 'Gen.Hunter' for Beaufort and returned at 4P.M. Left Port Royal for
Fortress Monroe July 1st and passed Charleston about 4 P.M. the same day.

Output

Request

```
▸ Object
  ▸ inputs: Object
  ▸ highlighter: Object
  image: "Base64 encoded image"
```

Response

```
▸ Object
  ▸ output: Object
  ▸ highlighters: Array [10]
```

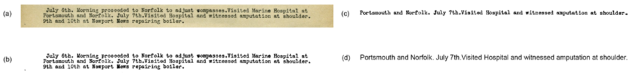
Figure 6 Results of a text line segmentation method. User input (left) is shown together with the computed results (right). Below the images is the JSON information a programmer would receive when calling the methods on DivaServices directly.

Figure 6 provides an example of a detailed result. The user input is shown (left) with the computed result (right). The image view can be manipulated (dragging, and zooming) to get a better view of certain areas. Below each image is the JSON information that is sent to and received from DivaServices. This information should help programmers to see with what kind of information they have to deal with should they decide to integrate that method into another application.

Using DivaServices-Spotlight for Designing DIA Workflows

We provide an example how DivaServices-Spotlight can be used to design a full workflow. The aim is to build a system that takes an input image and performs OCR on the segmented text lines. For this we need to perform three steps: binarization, text line segmentation, and OCR recognition.

Using the “Save Image” functionality on the result page we save the result image after each step. Figure 7 (a) – (d) show results at each stage using a combination of available methods. Parameters or even method could be changed at each step in order to find the best suited combination.


(a)
July 6th. Morning proceeded to Norfolk to adjust compasses.Visited Marine Hospital at
Portsmouth and Norfolk. July 7th.Visited Hospital and witnessed amputation at shoulder.
9th and 10th at Newport News repairing boiler.

(b)
July 6th. Morning proceeded to Norfolk to adjust compasses.Visited Marine Hospital at
Portsmouth and Norfolk. July 7th.Visited Hospital and witnessed amputation at shoulder.
9th and 10th at Newport News repairing boiler.

(c)
Portsmouth and Norfolk. July 7th.Visited Hospital and witnessed amputation at shoulder.

(d)
Portsmouth and Norfolk. July 7th.Visited Hospital and witnessed amputation at shoulder.

Figure 7 Results at different stages in the workflow. The input image (a) is binarized (b), segmented into text lines (c) and processed using an OCR algorithm, leading to its digital representation (d).

Once a researcher is satisfied with the results on a small scale, he could then integrate that workflow into his application by directly invoking the methods on DivaServices using his programming language of choice.

Conclusion

With DivaServices-Spotlight we provide a web application to interact with all available methods hosted on DivaServices. Researchers can run small scale experiments to experience the possibilities of the different algorithms. Furthermore, the application provides developers with the necessary information they would need to use the methods outside of DivaServices-Spotlight and integrate them into other applications.

Bibliography

Clausner, C., Pletschacher, S., and Antonacopoulos, A. (2011). Aletheia – An advanced document layout and text ground-truthing system for production environments. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 48–52.

Embach, M., Krause, C., Moulin, C., Rapp, A., Rindone, F., Stotzka, R., … Vanscheidt, P. (2013). eCodicology-Algorithms for the Automatic Tagging of Medieval Manuscripts. *The Linked TEI: Text Encoding in the Web*, pp. 172.

Würsch, M., Ingold, R., and Liwicki, M. (2015). DIVAServices – A RESTful Web Service for Document Image Analysis Methods. In *Digital Humanities.*

Notes

[1] Available at: https://github.com/DIVA-DIA/DIVAServices-Spotlight

[2] Available at: http://divaservices.unifr.ch/spotlight

Visualising the Dynamics of Character Networks

Aris Xanthos
aris.xanthos@unil.ch
University of Lausanne, Switzerland

Isaac Pante
isaac.pante@unil.ch
University of Lausanne, Switzerland

Yannick Rochat
yannick.rochat@epfl.ch
Swiss Federal Institute of Technology in Lausanne, Switzerland

Martin Grandjean
martin.grandjean@unil.ch
University of Lausanne, Switzerland

Introduction

The character network of a given narrative (novel, play, film, graphic novel, etc.) models the structure formed by the relations in its character-system (Woloch, 2003). A relation between two characters symbolises their co-presence in parts of the narrative; the entire set of relations between all characters constitutes a formal model of this character-system and lends itself to display and analysis. For example, Moretti (2011) used network modelling to compare the importance of protagonists from Shakespeare's *Hamlet*, while Trilcke *et al.* (2015) created character networks for 465 German plays and used them to initiate a wider study of German Theatre.

Most applications of character network analysis have disregarded temporality, possibly because of its representational complexity. Consequently, all relations in the system are considered as happening at the same time: one cannot distinguish if a given edge symbolises a relation at the start, at the end, or in several parts of the work under study. Furthermore, because temporality is not being accounted for, there is usually no way of relating the network visualisation with the unfolding of the source narrative. While prototypes such as those discussed in Roberts-Smith *et al.* (2013) offer sophisticate ways of dynamically visualising the text of theatre plays, they do so in a way that is unrelated to character network modelling.

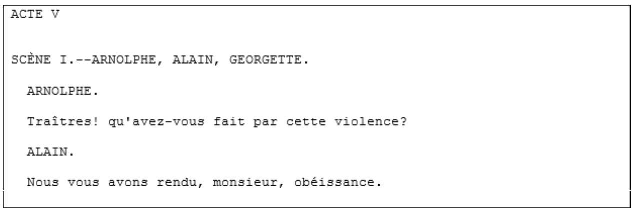
Based on these observations, we set out to develop an open source web application which models the character-system of theatre plays as a sequence of network states synchronised with the actual narrative content (https://github.com/maladesimaginaires/intnetviz). This paper proposes a high-level overview of our application , successively focusing on the underlying structure extraction process, the conception of the graphical interface, and the range of uses envisioned for it. In the conclusion, we evoke the ways in which we intend to develop it and reflect on the potential significance of this development at a more epistemological level.

Tool overview

The underlying workflow has been divided into two parts. First, the text of a play is processed using *Orange Textable* (http://langtech.ch/textable, Xanthos 2014), an open source text analysis add-on for the *Orange Canvas* (http://orange.biolab.si/) visual programming environment: in particular, the play is divided into its component parts (acts, scenes, lines) and the characters present in each scene are identified and associated with each line. These data are then imported into a web interface based on the open source *D3* JavaScript library (https://github.com/mbostock/d3, Bostock et al., 2011), which allows the user (author, researcher, teacher, etc.) to manipulate the character network without prerequisite installation.

Structure extraction

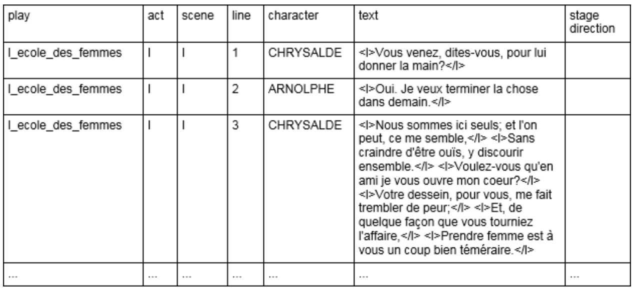
The workflow is designed to facilitate the later inclusion of new data. The play used in the initial development phase was Molière's *L'Ecole des femmes*, as found in raw text format on the Project Gutenberg website (http://www.gutenberg.org/files/43535/43535-0.txt). Theatre plays constitute a privileged input for the automatic extraction of structural information: such information is in general explicitly encoded in these data, as illustrated by the excerpt of Molière's play reproduced on Figure 1. It is worth noting that the extraction phase can be readily adapted to take advantage of cases where structural information has been formally encoded using TEI-XML annotation for instance, such as the "Théâtre classique" database used by Karsdorp *et al.* (2015). We are currently investigating the possibility of extending our approach to a large body of plays in this way.



```
ACTE V


SCÈNE I.--ARNOLPHE, ALAIN, GEORGETTE.

  ARNOLPHE.

  Traîtres! qu'avez-vous fait par cette violence?

  ALAIN.

  Nous vous avons rendu, monsieur, obéissance.
```

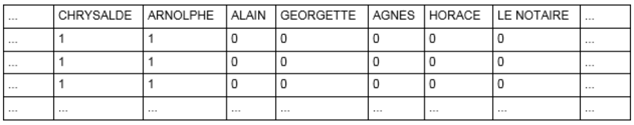
Figure 1. Excerpt of L'Ecole des femmes illustrating the explicit structuration of the data

Structure extraction is performed by a mostly linear chaining of segmentation and recoding operations based on regular expressions which gradually transform the raw text of the play into the tables that will be later used for controlling the web interface. Each table has the same number of rows, corresponding to the extracted lines of the play. The first table gives the text of each line (lightly annotated in XML) along with the associated character and stage directions (Table 1). The second table indicates the presence or absence of each character when each line is spoken (see Table 2).



play	act	scene	line	character	text	stage direction
l_ecole_des_femmes	I	I	1	CHRYSALDE	<l>Vous venez, dites-vous, pour lui donner la main?</l>	
l_ecole_des_femmes	I	I	2	ARNOLPHE	<l>Oui. Je veux terminer la chose dans demain.</l>	
l_ecole_des_femmes	I	I	3	CHRYSALDE	<l>Nous sommes ici seuls; et l'on peut, ce me semble,</l> <l>Sans craindre d'être ouïs, y discourir ensemble.</l> <l>Voulez-vous qu'en ami je vous ouvre mon coeur?</l> <l>Votre dessein, pour vous, me fait trembler de peur;</l> <l>Et, de quelque façon que vous tourniez l'affaire,</l> <l>Prendre femme est à vous un coup bien téméraire.</l>	
...	...	...	...	...	...	...

Table 1. First rows of the table containing the extracted text and stage directions



...	CHRYSALDE	ARNOLPHE	ALAIN	GEORGETTE	AGNES	HORACE	LE NOTAIRE	...
...	1	1	0	0	0	0	0	...
...	1	1	0	0	0	0	0	...
...	1	1	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...

Table 2. First rows of the table indicating the presence or absence of each character when each line is spoken (the first four columns are identical to Table 1 and have been omitted, along with the last few character columns)

Applied to *L'Ecole des femmes*, the process was found to give fairly reliable results: about 98% of the lines were correctly extracted. Most errors could be fixed by determining *a priori* the set of acceptable character names in a given play. An unexpected source of errors was the assumption that characters could be consistently identified by a single string; even without taking into account situations where a character is explicitly "renamed" as part of a surprise effect, linguistic processes such as the succession of indefinite and definite articles in French discourse lead to formal variations in the designation of characters.

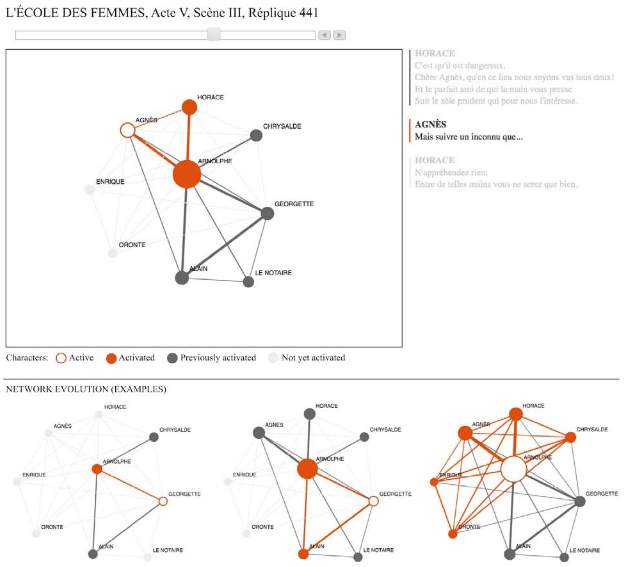
Interface

Data files are then imported into a browser and D3 is used to build an interactive visualisation of the character-system (a demo is available at http://bit.ly/network-demo) where the state of the network is synchronised with the current line (Figure 2). Character nodes are positioned using a force-based layout. Browsing the text simultaneously updates the network, line, and position indicator (act, scene, line). At each step, the weight of edges (expressed by their width) increases with the number of co-presences between the characters up to this point. Similarly, the weight of nodes (their radius) increases with the number of lines spoken by this character.

A character node can be in one of four states. It is:

- "active" when the character is speaking.
- "activated" when it is present in the scene but not active.
- "previously activated" if it was present in the play but is not currently active nor activated.
- "not yet activated" if its first appearance is in a later scene.

The latter state makes it possible to offer a view of the final network state, highlighting absences as well as presences in the earlier stages.



L'ÉCOLE DES FEMMES, Acte V, Scène III, Réplique 441

HORACE
C'est qu'il est dangereux,
Chère Agnès, qu'en ce lieu nous soyons vus tous deux!
Et le parfait ami de qui la main vous presse
Suit le zèle prudent qui pour nous l'intéresse.

AGNÈS
Mais suivre un inconnu que...

HORACE
N'appréhendez rien:
Entre de telles mains vous ne serez que bien.

HORACE
AGNÈS
CHRYSALDE
ARNOLPHE
ENRIQUE
GEORGETTE
ORONTE
ALAIN
LE NOTAIRE
Characters: ○ Active ● Activated ● Previously activated ● Not yet activated
NETWORK EVOLUTION (EXAMPLES)

Figure 2. User interface. [Above] Agnès is speaking in presence of Horace and Arnolphe; all characters have already intervened except Oronte and Enrique, who will appear in the final scenes. [Below] Three moments of the unfolding character network (lines 41, 359 and 536)

Uses

We display each line in front of the matching network state for a reason: the network does not exhaust the richness of the play. Since co-presence analysis does not consider the actual content of lines, it cannot–nor aims to–account for discursive references to characters. However, by extracting an actantial model from speech turns, the tool provides a concise and dynamic representation of the enunciative structure of the play. Paired with background knowledge about theatrical narration, such a visualisation offers new ways of reading.

From a philological perspective, relating sociological variables (gender, age, social status) with structural properties of the network (dynamical statistics, centrality measures, word and line counts) helps clarifying which profiles lead the interactions, which ones merely react to them, and which ones are excluded from them, revealing potential social stigma. By expliciting not only the presence but also the absence of links between certain characters, the proposed visualisation may contribute to a visual and interactive deconstruction of the plays (Derrida, 1967).

By allowing the user to browse through successive pictures of the interactions, the interface provides a unique opportunity to "play" the play and visualise the flow of speech between characters. We believe that this playful dimension is particularly interesting for pedagogic uses, especially when discovering a new narrative work with students.

Last but not least, our method can also be applied to any text in the course of the writing process. In this context, disposing of a visualisation of existing interactions at each moment of the text may help the writer distance herself from an impressionistic representation.

Perspectives and conclusion

One of the most significant benefits of a digital humanities approach is the dialectic relation created between the development of a prototype and the verification of scientific hypotheses. Discussing seemingly trivial design issues like the number of colors requested to map the network relations has frequently led us to expliciting and fruitfully questioning differences in our epistemological backgrounds. Thus we consider the following perspectives not only as potential improvements to our tool but also, more importantly, as opportunities to challenge our own theoretical preconceptions:

- Further enriching the visualisation (e.g. by adding line count histograms or highlighting of the first intervention of a character) would benefit the back-and-forth movement between the content and the structure of a given play.
- The dynamical nature of the interface would enable us to compare the distinctive features of different plays not only in terms of the final state of character networks, but also and more importantly, of their evolution; this should help bring out "interactional styles" and discover *Familienähnlichkeit* (Wittgenstein, 1953) by author or period.
- Other kinds of texts will benefit from such a dynamical analysis: other fictional texts such as screenplays, but also linguistic transcriptions as used in conversation analysis.

By allowing the user to browse the content of a narrative and manipulate an interactive character network, our motivation is ultimately to contribute to a better integration of distant and close reading practices.

Bibliography

Bostock, M., Ogievetsky, V. and Heer, J. (2011). D3: Data-Driven Documents. *IEEE Trans. Visualization and Comp. Graphics (Proc. InfoVis)*. http://vis.stanford.edu/papers/d3 (accessed Oct. 28, 2015.)

Derrida, J. (1967). *Of Grammatology*. Baltimore: The Johns Hopkins University Press.

Karsdorp, F., Kestemont, M. Schöch, C. and Van Den Bosch, A. (2015). The love equation: Computational modeling of romantic relationships in French classical drama. In *Proceedings of the 6th Workshop on Computational Models of Narrative* (CMN-2015), pp. 89–107.

Moretti, F. (2011). Network theory, plot analysis. *New Left Review*, **68**: 80–102.

Roberts-Smith, J., DeSouza-Coelho, S., Dobson, T., Gabriele, S., Rodriguez-Arenas, O., Ruecker, S., Sinclair, S., Akong, A., Bouchard, M., Hong, M., Jakacki, D., Lam, D., Kovacs, A., Northam, L. and So, D. (2013). Visualizing theatrical text: from Watching the Script to the Simulated Environment for Theatre (SET). *Digital Humanities Quarterly*, 7(3).

Trilcke, P., Fischer, F. and Kampkaspar, D. (2015). Digital Network Analysis of Dramatic Texts. *Digital Humanities 2015*. http://dh2015.org/abstracts/xml/FISCHER_Frank_Digital_Network_Analysis_of_Dramati/FISCHER_Frank_Digital_Network_Analysis_of_Dramatic_Text.html (accessed Oct. 28, 2015).

Wittgenstein, L. (1953). *Philosophical Investigations.* Blackwell Publishing.

Woloch, A. (2003). *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel.* Princeton University Press.

Xanthos, A. (2014). Textable: programmation visuelle pour l'analyse de données textuelles. *Actes des 12èmes Journées internationales d'analyse statistique des données textuelles (JADT 2014)*, pp. 691–703.

Work In A Globalised World. Allocation Algorithm To Add Labour Relations To Digitised Census Data

Richard Zijdeman
richard.zijdeman@iisg.nl
International Institute of Social History

Rombert Stapel
rombert.stapel@iisg.nl
International Institute of Social History

Contemporary studies on social inequality often focus on income or status attainment, but neglect the fundamental underlying relationship between employer and employee, referred to as a 'labour relation'. This negligence of labour relations (for and with whom one works) was for a long time of no concern, as the employer-employee relationship was predominant in the period after World War II. However, the current fierce debate on the desirability of precarious work among the rising number of freelancers and independent contractors around the world, as well as the increased media attention and consumer awareness of 'forced labour', shows that employer-employee relationship is not a universal constant, but the result of a much broader historical context. Moreover the interest in the historical context of social inequality is now bigger than ever, judging by both the academic and media impact of recent publications (e.g. Clark, 2014; Piketty and Goldhammer, 2014; van Zanden et al., 2014).

In order to describe and explain the historical context of shifts in labour relations and to recognize global connections between these shifts, a taxonomy of labour relations covering the past five centuries was devised by the participants of the Global Collaboratory on the History of Labour Relations (https://collab.iisg.nl/web/LabourRelations/; Figure 1). This collaboratory – hosted by the International Institute of Social History in Amsterdam (http://www.socialhistory.org) – has grown in the past decade to an online scholarly community of dozens of regional experts across the world and from various scholarly disciplines, such as social and economic historians, archaeologists and sociologists.

Using the taxonomy (Lucassen, 2013) (Figure 1), accompanying codebook and guidelines for data entry, the global collaboratory has gathered data on labour relations for more than twenty countries, using five temporal cross-sections (1500, 1650, 1800, 1900, [Africa: 1950], 2000) (Brown and van der Linden, 2010; Hofmeester and Moll-Murata, 2011; van der Linden, 2008; Lucassen, 2008). Databases of many more countries and regions are currently being prepared. For each country and cross-section a set of scholarly products are created that consists of a predesigned database with 1) population and demographic data and 2) details of the labour relations. A methodological paper explaining the choices made in the data collection for each country and time period accompanies the database. Both the database and methodological paper are verified before they are provided online as open access publication.

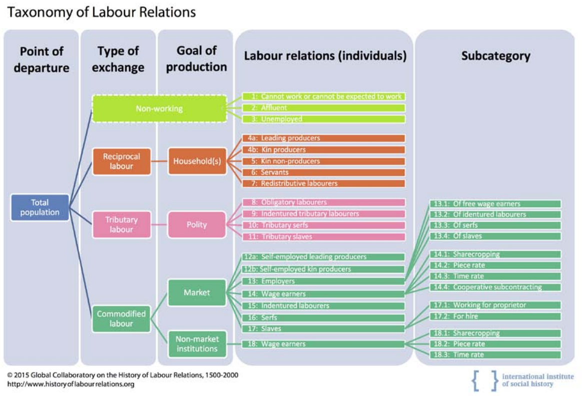


Figure 1. The taxonomy of labour relations, 2015

Until now, most of the data gathered are manually derived from aggregated sources, similar to contemporary occupational censuses, mainly for the period before 1800. The main aim of this paper is to present an alghorithm that was developed to automatically derive labour relations from digitized census materials. We therewith improve on the traditional way of working in the following ways. First, our alghorithm specifically applies to the biggest two projects that digitize census materials in the world NAPP and IPUMS, therewith providing a major contribution to the collection of labour relations for the post 1800 period and for the entire world.

A second advantage of the algorithm is that it provides the first alternative derivation of labour relations and thus can be used to test the reliability of the traditional derivation of labour relations. For many of the digitized censuses both original aggregated tables as well as the individual level data have been preserved, thus allowing for a reliability test of both methods.

The third advantage our algorithm provides is that is able to derive labour relations from individual level census data, thus allowing for more rigorous tests of hypotheses on labour relations. For until now, labour relations have just been gathered from aggregated census tables, only allowing for national comparisons and hypothesis testing. Also labour relations for the past two centuries have only been gathered for three cross-sectional years, while census data is available by decade from ca. 1850 onwards. By being able to attach labour relations to individuals, for the first time descriptive results on heterogeneity of labour relations within countries as well as over time will become available. Moreover, researchers will be much more able to zoom in on the characteristics related to changes in labour relations, such as individual characteristics (e.g. age, gender, education), household composition (e.g. extended family, sibling composition), and historical context of the municipality, region or state (e.g. level of development, political orientation) and therefore be able to provide more rigorous tests of hypotheses on changes in labour relations.

For this purpose, an algorithm is created to allocate labour relations to each of the millions of individual records that are part of the IPUMS databases. This algorithm was first tested on the censuses of the United States between 1850 and 2010. While an earlier version was written in R, the current algorithm was written in SPSS syntax, a fourth generation programming language, as SPSS proved to be better equipped for our purpose and to handle the large file sizes (4-10 GB). In the end, the algorithm will be available to all users of the IPUMS databases.

Starting from the total population, in each iterance a proportion of the records was allocated to a certain labour relation until all records were assigned. Different variables were used or combined, including age, class of worker, employment status, whether someone was considered in the labour force, their occupation, whether they lived on a farm. Here, the order of the different allocations is very important. Broadly speaking, the algorithm moves from the general to the detail. The end result is an enormous database of labour relations, both in the United States and the wider world. This allows us to study shifts in labour relations in much more detail than before (for example: Figure 2). As the census also includes numerous other information, including for instance place of residence, it is also possible to study in great detail geographical factors

(Figure 3, or the role of education, gender, age, ethnicity, household composition, migrant status, wealth, and many other things.

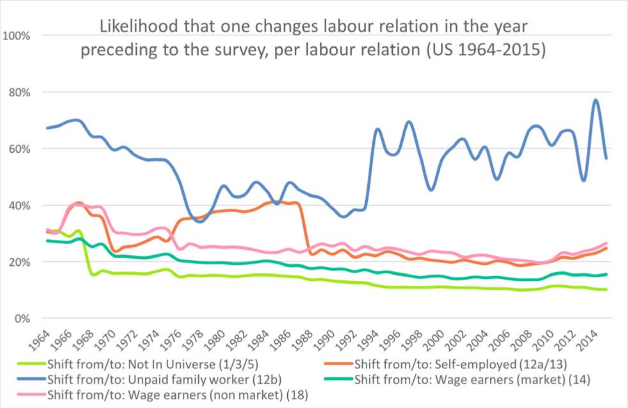


Figure 2. Likelihood that one changes labour relation

Although the IPUMS project has done much effort already to harmonise the different census data, and although historical census takers were also much aware of the need to create uniform censuses both national and trans-national, the algorithm can easily be adapted to the specifics of each of the sources used by the Collaboratory. This means that changes in categories in the census, but also changes in the meaning of census category labels can be adapted to. A future goal is to create an updated version of the algorithm, that allows not only an allocation of a labour relation, but also provides a certainty value. This update is expected in the coming months. Also, in our paper, in addition to presenting the algorithm itself, we will provide a complete research cycle for the US census data from 1850-2013. We start by deriving the labour relations for each of the census years and show our census-specific adaptations of the algorithm to account for historic changes in census taking, such as changes in instructions or the meaning of census category labels. Next, we will describes shifts in labour relations in the US for the period under study using amongst others animations as depicted in Figure 3. Finally, we will derive and test hypotheses on acts that affect self-employment in the US, such as the Midwives Act, that forced women out of self-employment.

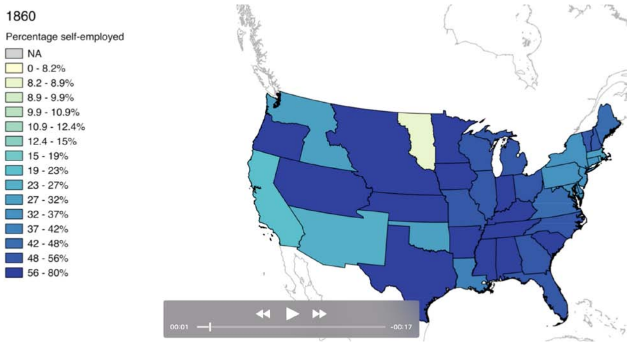


Figure 3. Still of an animation of changes in self-employment in the US, 1850-2013

Bibliography

Brown, C. and Linden, M. van der. (2010). Shifting Boundaries between Free and Unfree Labor: Introduction. *International Labor and Working-Class History*, (78), pp.4–11.

Clark, G. (2014). *The Son Also Rises.* [online] Princeton University Press. Available at: http://www.jstor.org/stable/j.ctt5hhrkm.

Hofmeester, K. and Moll-Murata, C. (Eds.). (2011). *The joy and pain of work: global attitudes and valuations, 1500-1650*. International review of social history. Cambridge: [Published for the Internationaal Instituut voor Sociale Geschiedenis] Cambridge University Press.

Linden, M. van der. (2008). *Workers of the world: essays toward a global labor history*. Studies in global social history. Leiden ; Boston: Brill.

Lucassen, J. (Ed.). (2008). *Global labour history*. International and comparative social history. Bern ; New York: Peter Lang.

Lucassen, J. (2013). Outlines of a history of labour. (51), pp.5–46.

Piketty, T. (2014). *Capital in the Twenty-First Century*. [online] Harvard University Press. Available at: <http://www.jstor.org/stable/j.ctt6wpqbc>.

Van Zanden, J.L., Baten, J., Mira d'Ercole, M., Rijpma, A. and Timmer, M. (Eds.). (2014). *How was life? Global Well-being since 1820*. OECD Publishing.

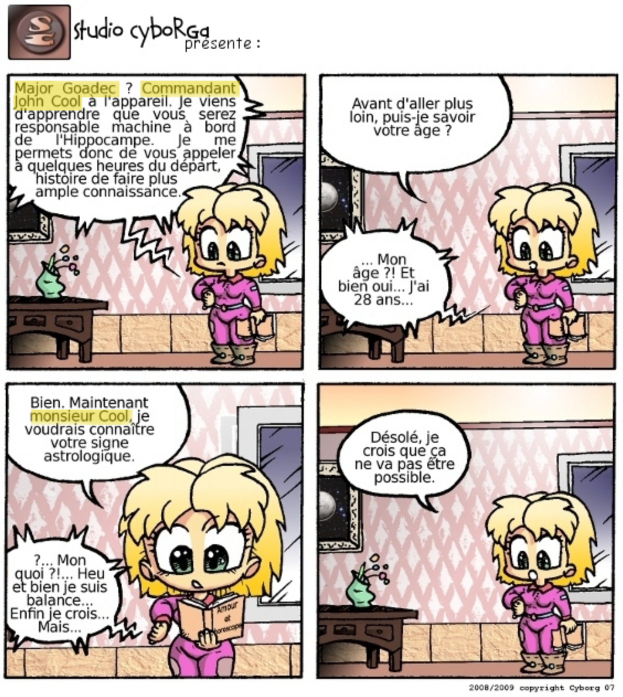
Short papers

Automatic Detection of Characters in Case Insensitive Text in Comics

Alaa Abi Haidar
alahay@gmail.com
LIP6. Univeristy of Pierre and Marie Curie (UPMC), France

Jean-Gabriel Ganascia
jean-gabriel.ganascia@lip6.fr
LIP6. Univeristy of Pierre and Marie Curie (UPMC), France

Huge amounts of comics are being published on a daily basis. However, the textual data that are circumscribed in comics' speech balloons and thought bubbles are highly unstructured, and very hard to automatically extract and digitize. Several automatic techniques have been used to automatically detect the shapes, sizes, and contents of speech balloons and thought bubbles leading to automatic text digitization and localisation in comics (Rigaud et al 2013; Ho et al 2012). Such advances open new horizons for a myriad of text mining applications to comics, namely, automatic indexing, searching, recommendation and visualization.


Studio cyboRGa présente :
Major Goadec ? Commandant John Cool à l'appareil. Je viens d'apprendre que vous serez responsable machine à bord de l'Hippocampe. Je me permets donc de vous appeler à quelques heures du départ, histoire de faire plus ample connaissance.
Avant d'aller plus loin, puis-je savoir votre âge ?
... Mon âge ?! Et bien oui... J'ai 28 ans...
Bien. Maintenant monsieur Cool, je voudrais connaître votre signe astrologique.
?... Mon quoi ?!... Heu et bien je suis balance... Enfin je crois... Mais...
Amour et horoscope
Désolé, je crois que ça ne va pas être possible.
2008/2009 copyright Cyborg 07

Named Entity Recognition (NER) is a task of information extraction under text mining that aims to identify in-text references to concepts such as people, locations and organizations, mainly in unstructured natural-language text. NER is very useful for text indexing, text summarization, question answering and several other tasks that enhance the experience between humans and literature.

In a previous study, we developed a simple and original method of Unsupervised Named Entity Recognition and Disambiguation (UNERD) (Mosallem et al 2014) to automatically extract names of people, locations and organizations from French and English (Abi Haidar et al., 2016) newspapers. We then used the text coordinates in the ALTO XML format to automatically locate and highlight the named entities detected by UNERD on the scanned image of the newspaper (Abi Haidar et al 2014). Last, we used UNERD to detect and visualize named entities from French newspaper during the period of the first world war (Abi Haidar et al 2016).

The main challenges encountered in unsupervised NER have been identified and addressed with our original UNERD method when applied on English and French newspapers and French literature. These challenges include but are not limited to named entity disambiguation, named entity boundary detection, and domain-specific dictionary attribution. In comics, in addition to the aforementioned challenges, the case-insensitive text extracted from comics adds the challenge of proper noun detection that we used to handle using POS tagging. We used TreeTagger (Shmid 2005) POS tagger to detect proper nouns, however and like other POS taggers, it works only when applied on case sensitive text.

Here, we use a variation of our UNERD method to detect names of fictional characters in a database of unstructured text from comics that has been recently digitized by our partners at the L3i Labs using an active contour model for speech balloon detection (Rigaud et al 2013). Au lieu of POS tagging for proper noun extraction, we filter stop words that constitute a 5000 word list of most frequent French words. For the dictionary, we use the freebase dictionaries of fictional characters and characters in fiction novels from Freebase. The comics digitised data amounts to around 8000 case-insensitive words in the French language. We evaluate our method's precision by analysing the predicted character names.

Out of 182 predictions of character names made by UNERD, 113 names were correctly predicted. This precision of 62% cannot be compared to that obtained in our previous results with UNERD when tested on case sensitive textual data. We also did not compute the coverage since we do not have any gold standard or any annotated data. In the table below, we see the most frequent characters that were automatically detected. The terms 'FLIP' and 'HUM' could be discarded by adding a list of onomatopoeic terms to the list of stop words thus improving precision to 68%. Our precision compares well to results presented in Cornolti's meta-study of Named Entity Recognition/ Entity Linking resulting in a precision of 69% (Cornolti et al 2013).The extracted character names are then associated with page numbers to help with the automatic indexing of characters that is otherwise very costly in time and human resources. Co-occurrences of character names in the same page are used to detect dialogues or interactions between the characters. Such information, along with the mere occurrences of character names, are used by our partner, Actialuna, in order to enhance digital comics recommendation. Our method can be applied to extract entities from other case-insensitive texts by simply adapting the dictionary to the targeted text. Our results are preliminary and have been tested on a tiny database that is currently being updated.

Frequency	Entity
2	CASTOR
2	COW-BOY
2	CYBORG
2	Doc
2	FRED
2	Keuf
2	Lamisseb
2	LET
2	LONGUE-VUE
2	PHILÉMON
2	PiLOT-BOAT
2	PRESIDENT
2	RAT
2	Stock
3	HUM
3	JEAN-MICHEL
3	JOHN
3	KID
3	LEO
3	Li
3	MIN
3	NEAR
3	NEW-YORK
3	PiLOTE
3	PIRANHA
3	Traffic
4	ACHETé
4	BILL
4	DEMON
4	DOC
4	ZiG
5	MAN
5	STEVE
8	BOY
8	CINDY
8	DOLLY
8	FLIP
12	NEMO
14	COCO
14	SUPER
16	ALFRED

Bibliography

Rigaud, Ch., et al. (2013). An active contour model for speech balloon detection in comics. *Document Analysis and Recognition (ICDAR)*, 2013 12th International Conference on IEEE.

Ho, A. K. N., Burie, J.-Ch. and Ogier, J.-M. (2012). Panel and speech balloon extraction from comic books. *Document Analysis Systems (DAS)*, 2012 10th IAPR International Workshop on IEEE.

Mosallem, Y., Abi-Haidar, A. and Ganascia, J. G. (2014). Unsupervised Named Entity Recognition and Disambiguation: An Application to Old French Journals. *Proceedings of ICDM 2014*. St. Petersburg, Russia.

Alaa ABI HAIDAR, Jean-Gabriel GANASCIA (2014). Mapping French Press to the Digital Age. at Digital Humanities 2014 Conference. DH 2014: 7-12 July 2014. Lausanne, Switzerland.

Alaa ABI HAIDAR, Oscar ALBERTINI, Jean-Gabriel GANASCIA (in press). A Simple yet Efficient Unsupervised Named Entity Recognition Model. Wiley's DMKD (In Press).

Schmid, Helmut (1995). Treetagger, a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, **43**: 28.

Cornolti, Marco, Paolo Ferragina, and Massimiliano Ciaramita (2013). A framework for benchmarking entity-annotation systems. *Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee*, 2013.

Alaa Abi-Haidar, Bin Yang, Jean-Gabriel Ganascia (2016). Mapping the First World War Using Interactive Streamgraphs. *Sociology and Anthropology*, **4**: 12-16.

Minimal Editions in the Classroom: A Pedagogical Proposal

Susanna Allés Torrent
susannalles@gmail.com
University of Miami, United States of America

Alex Gil
ag3339@columbia.edu
Columbia University, United States of America

Introduction

Mastering a complete workflow for creating editions when the editor works alone, or has no access to technologists, has proven challenging, leading to the creation and adoption of tools that make the final product conform to pre-determined models (Burnard et al., 2006; Pierazzo, 2015). Although having access to such tools can be very salutary in many cases, the need remains for independent scholars to be able to create their own editions based on unique visions, without recourse to grants or large teams. One of the main challenges we face to achieve these goals is our limited ability to train humanities students in the span of a semester the full stack of skills they would need in order to create digital editions according to standards (MLA, 2011). We believe that by adopting a minimal computing (http://go-dh.github.io/mincomp/) approach, we can achieve such a goal.

This paper reports on a classroom experiment done during a course entitled *Creating a Minimal Edition: From the Manuscript to the Web*, that tries to address these issues in context. The main goal of the course is to introduce students to textual scholarship in general, and to digital scholarly editing in particular (Fraistat & Flanders, 2013). Our core goal is to build a TEI-to-Jekyll workflow and a simple customizable web template to be used for small digital editions, and release it openly to scholars. Underneath our prototype, we adopt the principles of minimal computing. By minimal we understand computing done under a set of significant constraints, where we reduce the stack to the minimum needed technology to accomplish a scholarly need: in this case to produce a critical digital edition that meets standards and teach the necessary skills to students within the course of one semester.

We use our created template to publish a small-scale digital scholarly edition online of one of the most remarkable Spanish literary works, the *Lazarillo de Tormes* (16th century). The course is conceived as a combination between collaborative research and digital humanities design. At all steps of the process, instructors and students work together toward the completion of the digital edition. The course is divided into lectures and recitation sessions. The first is an opportunity for students to become familiar with central ideas in textual scholarship, while focusing on the textual situation of the Lazarillo; the second allow us to put those ideas into practice in the creation of the digital edition.

Underneath our proposal, there are three main theoretical axes of different nature: ethical, technological, and scholarly.

First, we would like to underscore some issues related to execution and interfaces. Well funded research projects may afford IT services or the creation of standalone systems or "workstations" to build and publish their edition, but this is the best case scenario which sometimes is made publicly available; but not always these interfaces are flexible enough and not usable by people with few IT knowledge. Digital scholarly edition, and specially workflow and interfaces, need to be more diffused, more used, easier to build and less expensive, inspired in what has been called the "bricks approach" (Pierazzo, 2015: 116-117). Hence, the need to offer a simple framework that begins on an XML-TEI dataset and ends in a minimal interface, customizable by the user.

Second, and tightly connected with the ethical component, is the relevance of technological choices, which deals with two main issues: the use of available standards and the open source ethos. We believe that digital scholarly edition, and DH in general, must rely on free and standards web services and technologies. Furthermore, we align our project on what it has been called "open source critical edition" (Bodard and Garcés, 2009) underscoring the need to make available all the datasets and scripts involved in the project.

Third, we are aware that for our show-case scenario the term "scholarly edition" is somewhat simplified. For our goal we understand "scholarly" in a large sense where the scholarly paradigm and a critical approach to primary sources are applied. Our work consists on the critical representation of one historical document, taking into consideration the full textual tradition (4 printed editions from the 16th century) of the literary work. The scholarly paradigm is obviously rethought from the digital perspective, giving special attention to execution and workflow, that is to say, giving room to understand concepts such as modeling and presentation. Furthermore, we wish to insist in two other main issues connected with DH discussions. On one hand, the fact that scholars need to take full control of their digital tools, as well as understand digital methods from their hermeneutical point of view. On the other one, we explore complex issues such as collaboration and authorship, in our case, wagering for a GitHub proposal.

Description of the Course

The course is divided into six main units and is conceived as a collaborative project, where each student is in charge of a main chapter of the literary work.

First, we offer a general introduction to textual scholarship and text editing, paying attention to scholarly editing trends from the 19th century to the present, through a selection of core readings in the field. We also offer a theoretical framework for digital editions, specifically to help students understand how digital editions differ from their traditional counterparts. Afterwards, students are introduced to Github and acquire the methodology needed to work collaboratively. The goal of these two sessions is to create a collaborative and robust work environment, and to ensure that all students gather the basic skills to become fully involved. In Unit 2, we present the primary source: the historical context, the argument, the literary relevance and the text of the *The Life of Lazarillo de Tormes and of His Fortunes and Adversities*. This work, published in 1554, is one of the first novellas in Iberian literature and a classic in the picaresque Spanish tradition. We then start planning the digital scholarly edition and its workflow. Students gather the basic approaches to the data modeling and conceptualize the text as a document object, starting from the analysis of the primary source. The next Unit is devoted to the eXtensible Markup Language and the Guidelines of the Text Encoding Initiative. Students gather the basic principles of the Extensible Markup Language, following the Guidelines of the Text Encoding Initiative. We also offer a general overview of the concepts of schemas (RelaxNG) and ODD documents. Each student is in charge of pursuing a textual encoding, marking up the main features: structural parts, typographic features, dates, place, and person names. They also become aware of the process of quality assurance of the text encoding of their peers.

The next steps of the process consists, on the one hand, in introducing students to the basics of Markdown, HTML and CSS, giving students the opportunity to think about data transformation, and to participate in the design and the final presentation format of the edition. On the other, we focus on inputs and outputs and textual migrations. The central node is the eXtensible Stylesheet Language Transformation, and the conversions from text encoding (XML-TEI) to the web (Markdown/HTML).

The last part of the course is devoted to web infrastructure and web publication. Students learn how to build a static website with Jekyll, dealing with the different technologies needed (HTML, CSS, Liquid, Markdown), and how to transfer their work from GitHub to GitHub Pages. We conclude with a minimal introduction to JavaScript, meant to introduce students to simple document interface: in this case the manipulation of the dates, places and person names marked up in the TEI.

As learning goals, we want our students to be able to participate in an authentic research and editing project, engaging in all steps of the process; to become aware of the challenges and opportunities of the digital medium for scholarly research and editing. We aim to teach how to apply different methods and technologies, to grasp the value of standards, and to understand data modeling and transformation from a theoretical as well as a practical perspective (Rehbein & Fritze, 2012). As a "technical" outcome, we seek to offer the basic skills to work independently in several languages (XML – TEI, HTML and CSS, XSLT, Markdown, Liquid, JavaScript), and a basic understanding of infrastructure (Jekyll, GitHub, Github Pages). Because this course is meant for students of Spanish as well, our program allows students to improve their Spanish language skills while engaging in public-facing, task-driven scholarship.

Our presentation will give us the opportunity to present the online version of the course, the results of the collaborative edition created along the 27 lectures of the semester, and, finally, to release the first prototype for minimal editions as a Jekyll template, in the hope that can be useful to other DH courses and projects.

Bibliography

Bodard, G., Garcés, J. (2009). Open Source Critical Editions: A Rationale. *Text Editing, Print and the Digital World*. Aldershot: Ashgate, pp. 83-98.

Burnard, L., O'Brien O'Keeffe, K. and Unsworth, J. (2006). *Electronic Textual Editing*. New York: Modern Language Association of America.

Fraistat, N., Flanders, J. (2013). *The Cambridge Companion to Textual Scholarship*. Cambridge: Cambridge University Press.

Modern Language Association (2011). *Guidelines for Editors of Scholarly Editions.*

Pierazzo, E. (2015). *Digital Scholarly Editing. Theories, Models and Methods*, Aldershot: Ashgate.

Rehbein, M. and Fritze, Ch. (2012). Hands-on Teaching Digital Humanities: A Didactic Analysis of a Summer School Course on Digital Editing. *Digital Humanities Pedagogy. Practices, Principles and Politics.* Open Book Publishers, pp. 47-78.

Testing the Doctrine of Election: A Computational Approach to Karl Barth's Church Dogmatics

Christopher Scott Bailey
csb5t@virginia.edu
Scholars' Lab, University of Virginia, United States of America

Eric Rochester
err8n@eservices.virginia.edu
Scholars' Lab, University of Virginia, United States of America

Karl Barth's *Church Dogmatics* (Barth, 1969-80) is widely considered to be one of the most influential works of Christian theology since the Reformation, and within it Barth's doctrine of election is considered a decisive contribution to modern theology (Webster, 2004: 1, 88, 93; von Balthasar, 1992: 174). Briefly, the doctrine of election, as a theological topic, describes the manner of God's salvific work, particularly how God determines those who will be saved. Barth's genius was to inscribe election within the relationship of the Father and the Son, Jesus of Nazareth, so that God is both the electing God and the elected human, and the humanity of Jesus is elected even while the divinity of Jesus is condemned. Over the past two decades, theologians have engaged in a rich questioning of the significance of Barth's doctrine of election for his own theology. His elaboration of the doctrine occurs in the document sections, traditionally referred to as paragraphs,[1] numbered 32 through 35 (of a total of 73 plus a fragment), with paragraph 33 being the primary location for Barth's innovative reworking of the doctrine of election. Some scholars, such as Bruce McCormack (McCormack, 2000, 2010) and Paul Jones (Jones, 2011), contest that election is a turning point in Barth's theology, decisively shaping the remainder of the *Dogmatics* to the point that some formulations after election are incompatible with formulations made prior to the doctrine of election, particularly within the context of Barth's Trinitarian ontology. Others, such as George Hunsinger (Hunsinger, 2000, 2008) and Paul Molnar (Molnar, 2002, 2006), argue that while the doctrine of election is the heart of the *Dogmatics*, it is a part of a consistent and coherent whole, and does not mark an incompatibility between what comes before and after.

This paper engages the question of the significance of the doctrine of election, as elaborated in paragraphs 32 through 35, to the whole of the *Church Dogmatics* through algorithmic approaches. It suggests that if a portion of a corpora strongly determines the rest of the corpora after its appearance, there will be textual traces, such as changes in word frequencies and common semantic groupings, that can be detected through computational analysis. It approaches the corpora, consisting of the entire *Church Dogmatics*, including prefaces and forewords written by Barth as well as his unfinished fragments that have been published as the final volume of the *Dogmatics*, though a variety of analytic techniques.

The initial explorations are conducted through topic modeling in order to discover hidden thematic structure in texts (Blei, 2012). Using Mallet, we first run topic models on different collections of paragraphs, from 15 to 30 topics, to discern the thematic structure of the entire corpora, noting especially those topics that seem definitively about the doctrine of election. Given the hypothesis that the *Dogmatics* from paragraph 36 on is determined by the theme of election in a way that the paragraphs leading up to paragraph 32 are not, we break the corpus into paragraphs 1 through 31, 32 through 35, and 36 through 73 plus the fragment that Barth was writing before his death. We then run topic models with the number of topics ranging from 15 to 30, looking for the presence of topics indicating the doctrine of election. We also run similar models for the entire corpus minus paragraphs 32 through 35 in order to see whether election would appear as a theme in the *Dogmatics* without the presence of the paragraphs explicitly committed to explicating the doctrine. Examining the results, we find that election fails to surface as a topic at most levels of granularity when paragraphs 32 through 35 are not included. We find that at all levels of granularity in which the topics are meaningful and coherent, election does not appear as a topic in the corpus consisting of paragraphs 36 through 73, plus the fragment, and we offer an interpretation for why this is the case based on the rhetorical strategy that Barth employs throughout his lengthy work.[2]

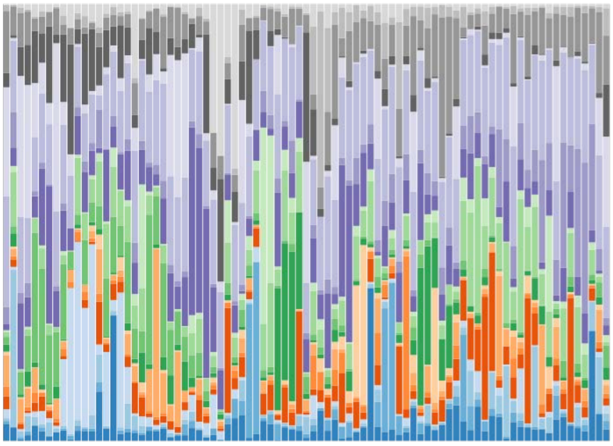


Fig. 1: Graph of topic distributions across the Church Dogmatics. Each column is one paragraph

Our topic models not only provide data for interpretation, but also supply a vocabulary for focusing further computational analysis. Based on words we determined to be distinctive to the theme of election, we examine overall frequency of key terms across the whole corpora, tracking the rise and fall of language specific to election. We also use term frequency-inverse document frequency (tf-idf) to examine which terms are particularly characteristic of individual paragraphs, paying attention to words typically associated with election (Kilgarriff, 2001; Garside, 2000). In a similar vein to that of our topic models, we determine mean tf-idf values for all features (words) in the broken down corpus, consisting of the same three chunks as determined above, at the levels of unigrams, bigrams, and ngrams (n=1-3). Looking especially at the bigram and ngram results, we do see shifts in the importance of some features that fit the hypothesis that Barth's doctrine of election determines the rest of his work. If the proponents of this thesis are correct, there should be an increase in reference to Jesus Christ due to a stronger Christological shift, and a greater sense of the humanity of Jesus due to the election-based eternal identification of the Son, the second Person of the Trinity, with the historical human Jesus of Nazareth. We do see a rise in the importance of 'jesus' in the unigram set, and in the ngram set see the bigram 'jesus christ' appear in the election and after election corpora. In the bigram set, we see 'elected man' appear in the election set, which in Barth's paragraph 33 references Jesus, and interestingly find "man jesus" as the fifth most characteristic bigram of the after election corpora. In relation to the text, we interpret this as an indication of the increased importance of the humanity of Christ in the election and after election corpora.

Based on these results, we explore two conclusions. While our analysis of frequencies and tf-idf values does seem to support the hypothesis that Barth's doctrine of election is a determining point in the *Dogmatics*, the connection between high value features and a substantial conceptual shift is difficult to determine, especially when the conceptual shift regards that ontology underlying theological developments in many doctrinal loci. Our topic models likewise were inconclusive in identifying shifts in the corpus that could be attributed to the paragraphs on election. We suggest that Barth's style of writing, which notoriously circles around and repetitively approaches topics from different angles, though with a traditional theological vocabulary, proves resistant to current algorithmic approaches in textual analysis.

	Before Election (1-31)		Election (32-35)		After Election (36-73)	
	feature	tfidf	feature	tfidf	feature	tfidf
0	god	0.45348	god	0.60445	god	0.38280
1	church	0.14875	man	0.22406	man	0.25072
2	note	0.14033	election	0.20015	jesus	0.10635
3	word	0.12710	jesus	0.16878	christ	0.08423
4	man	0.10055	israel	0.15690	does	0.07315
5	revelation	0.08683	christ	0.14954	note	0.05834
6	dogmatics	0.05937	note	0.13912	life	0.04063
7	christ	0.05479	divine	0.08037	command	0.04059
8	jesus	0.03732	doctrine	0.06774	christian	0.03982
9	father	0.03429	church	0.06342	world	0.03904
10	human	0.03273	elect	0.04152	community	0.03442
11	proclamation	0.03117	judas	0.03670	fact	0.03000
12	faith	0.02942	mercy	0.03296	divine	0.02823
13	son	0.02752	predestination	0.03212	human	0.02723
14	knowledge	0.02658	does	0.03122	faith	0.02602

Fig. 2: Top 15 tf-idf weights at unigram level

	Before Election (1-31)		Election (32-35)		After Election (36-73)	
	feature	tfidf	feature	tfidf	feature	tfidf
0	word god	0.10073	jesus christ	0.43971	jesus christ	0.15529
1	jesus christ	0.06273	doctrine election	0.06080	command god	0.03275
2	god word	0.03257	election jesus	0.05729	man woman	0.01668
3	knowledge god	0.02747	seven thousand	0.04377	new testament	0.01464
4	church proclamation	0.02747	community god	0.03803	man jesus	0.01428
5	new testament	0.01758	divine election	0.03569	divine command	0.01370
6	father son	0.01709	god mercy	0.03535	word god	0.01288
7	holy spirit	0.01542	god man	0.03517	god command	0.01253
8	doctrine trinity	0.01508	elected man	0.03418	soul body	0.01237
9	old testament	0.01484	election israel	0.03380	god man	0.01183
10	love god	0.01317	election grace	0.03314	holy spirit	0.01178
11	teaching church	0.01313	doctrine predestination	0.03123	theological ethics	0.00933
12	son spirit	0.01247	electing god	0.03121	holy day	0.00882
13	holy scripture	0.01231	new testament	0.03098	christian community	0.00867
14	love neighbour	0.01219	old testament	0.03036	christian love	0.00850

Fig. 3: Top 15 tf-idf weights at bigram level

	Before Election (1-31)		Election (32-35)		After Election (36-73)	
	feature	tfidf	feature	tfidf	feature	tfidf
0	god	0.36898	god	0.57432	god	0.32264
1	church	0.10521	man	0.21312	man	0.20533
2	word	0.09862	election	0.18994	jesus	0.08604
3	note	0.09168	jesus	0.16070	christ	0.06491
4	man	0.06653	israel	0.14977	does	0.04691
5	revelation	0.06392	christ	0.14209	note	0.04562
6	dogmatics	0.03703	note	0.13180	jesus christ	0.04511
7	christ	0.03666	divine	0.07615	command	0.03341
8	word god	0.03500	jesus christ	0.07115	world	0.03072
9	jesus	0.02710	doctrine	0.06378	christian	0.02614
10	father	0.02529	church	0.06062	life	0.02360
11	proclamation	0.02051	elect	0.03943	divine	0.02205
12	son	0.01820	judas	0.03485	community	0.02124
13	spirit	0.01721	mercy	0.03150	human	0.01906
14	holy	0.01623	predestination	0.03063	faith	0.01683

Fig. 4: Top 15 tf-idf weights at ngram(n=1-3) level

Bibliography

Balthasar, H. U. von. (1992). *The Theology of Karl Barth: Exposition and Interpretation*. Trans. Edward T. Oakes. San Francisco: Ignatius.

Barth, K. (1969-80). *Church Dogmatics*, 13 part volumes. Ed. G.W. Bromiley and T.F. Torrance. Edinburgh: T and T Clark.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, **55**(4): 77-84.

Hunsinger, G. (2000). *Disruptive Grace: Studies in the Theology of Karl Barth.* Grand Rapids: Eerdmans.

Hunsinger, G. (2008). Election and the Trinity: Twenty-Five Theses on the Theology of Karl Barth. *Modern Theology*, **24**(2): 179-98.

Jones, P. D. (2001). *The Humanity of Christ: Christology in Karl Barth's Church Dogmatics.* London: T and T Clark.

Jüngel, E. (2001). *God's Being is in Becoming: The Trinitarian Being of God in the Theology of Karl Barth. A Paraphrase.* Trans. John Webster. Edinburg: T and T Clark.

Kilgarriff, A. (2001). Comparing Corpora. *Journal of Corpus Linguistics*, **6**(1): 97-113.

McCormack, B. (2010). Election and the Trinity: Theses in response to George Hunsinger. *Scottish Journal of Theology*, **63**(2): 203-24.

McCormack, B. (2000). Grace and Being: The Role of God's Gracious Election in Karl Barth's Theological Ontology. In John Webster (ed.), *The Cambridge Companion to Karl Barth.* Cambridge: Cambridge University Press, pp. 92-110.

McCormack, B. (1995). *Karl Barth's Critically Realistic Dialectical Theology. Its Genesis and Development, 1909-1936.* Oxford: Clarendon Press.

Molnar, P. D. (2002). *Divine Freedom and the Doctrine of the Immanent Trinity: In Dialogue with Karl Barth and Contemporary Theology.* London: T and T Clark.

Molnar, P. D. (2006). The Trinity, Election, and God's Ontological Freedom: A Response to Kevin W. Hector. *International Journal of Systematic Theology*, **8**(3): 294-306.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh UP.

Rayson, P. and Garside, R. (2000). Comparing Corpora using Frequency Profiling. *Proceedings of the Workshop on Comparing Corpora*, **9**: 1-6.

Webster, J. (2004). *Barth*. London: Continuum.

Notes

[1] Barth's *Church Dogmatics* are broken down into volumes, part-volumes, paragraphs, and sections. Paragraphs are the primary organizing unit, though, and each constitutes a coherent thematic whole.

[2] In his text, *Barth*, Webster notes that Barth's, 'preferred method of exposition, especially in the *Church Dogmatics*, is frustrating for readers looking to follow a linear thread of argument. Commentators often note the musical structure of Barth's major writings: the announcement of a theme, and its further extension in a long series of developments and recapitulations, through which the reader is invited to consider the theme from a number of different angles and in a number of different relations" (13). Barth frequently circles around his topics, returning again and again to various doctrines through different paragraphs, in each case attempting to approach in such a way as to show something new.

Concept Modeling the Advertising Chinese Modern Society

Tani Barlow
barlow.tani@gmail.com
Rice University

Jing Chen
ccjj2008@gmail.com
Nanjing University

Ke Deng
kdeng@math.tsinghua.edu.cn
Tsinghua University

In the late 19th century, thousands of industrially produced consumer items flooded into extraterritorially governed, internationally regulated, Chinese, treaty port cities. Foreign commodities were products, and formed the backbone of new, urban, popular consumer culture. Consequently, the advertising industry infiltrated commodity brands and branding techniques into everyday life making commodity images a paramount symbol of civilized urban life. Advertising ephemera thus provides researchers with the conditions for thinking about modernity par excellent since it breaks data free of its origins to demonstrate how concepts embedded in ads ingratiate all consumer cultures (Barlow, 2012).

To force advertising to speak clearly, we launched the Chinese Commercial Advertisements Archive ("CCAA") and 'metadated' (Lev Manovich's term) more than ten thousand high quality images from microfilm copies of three, major, commercial, Chinese newspapers, in the period of 1880 to 1940 (Manovich, 2002). CCAA applies customized metadata schema based on the structural standard, Dublin Core, to each digital image of advertisement, entering all relevant information e.g., cartoon, brand icon, word texts and syntax, plus street names and business titles. Our metadata include: descriptive content, contextual information, bibliographical, technical and image sources of location, copyright status, and owning institution.

Scholars had already studied categories like hygienic/卫生), modern/现代, human/人类, eugenics /优生, and female/女性 in commercial/common ideas. They sampled image-based advertisements in libraries using newspapers, facsimiles and microfilm/fiche. Though more recent research projects have done a poor job of digitizing advertising, still we cannot ignore ad digitalization because historians are still generalizing from a fraction of ads that make up any potential archive. To avoid wasting time and to collaborate with other scholars developing what Franco Morreti calls 'distance reading,' we seek to connect concepts appearing in advertisements to concepts found in sociological texts employing statistical text mining of advertising copy (Hayles, 2012).

Space prevents a full literature survey here, but we have met with pioneering researchers Professor Peter Bol of 'China Biographical Database Project' at Harvard University and Professors Zheng Wenhui and Liu Zhaolin, co-PIs, 'Database for the Study of Modern Chinese Thought and Literature (1830–1930) 'at Taiwan National Chengchi University, and now have available over 30,000 annotated ad images which our proposed paper will use to augment evidences and expand analysis.

On the basis of these 30,000 annotated, newspaper, advertising images, our work is generating a text-mining model for advertising language, a language presenting historically anchored technical difficulties as follows:

1. Lack of word boundaries and punctuation. Word boundaries in Chinese are invisible; worse, ad slogans are not punctuated. Raw data is just a sequence of unsegmented Chinese characters which means text mining in Chinese is comparatively tough.

2. Lacks definitions of vocabulary. Ad texts contain lots of instable, idiosyncratic technical terms, like company names written in different ways, transliterated brand names, product names and so on that we discover during the text mining process.

3. Lack of training data. Most Chinese text mining methods depend on high quality training data, and will fail if the target texts are remarkably different from the training data. Considering that the advertisements that interest us are from regional newspapers over a long period, ad writing style is uncertain due to local linguistic differences. We cannot rely on current training data employing modern Chinese to establish models for mining 1920s syntax, vocabulary, punctuation (or lack of it) word use, semantic references, ideograph variation for 100-year-old print media.

4. Difficulties distinguishing technical and background words. Ad texts are a mixture of technical and background phrases, so it is not a trivial task to distinguish technical terms, our true interest , from noise, words rarely used a century after the ads were published.

We have overcome many of these roadblocks using statistical methods for Chinese text mining and knowledge discovery. Text mining allows us to: 1) discover potential associations among features and terms extracted from advertisements; 2) build links among these and ideological trends in the treaty port urban areas of China during our period by developing Deng Ke's statistical text mining method to establish indices of technical terms ("TT") and metadated association patterns among technical terms ("APTT") (Deng, Geng and Liu, 2014). Word Dictionary Model ("WDM") and Advanced Word Dictionary Model ("AWMD") are tools for word discovery, text segmentation and entity recognition of Chinese texts when training data are not available. WDM can be extended into an AWDM to achieve automatic recognition of TT (distinguishing technical terms from background words/phrases). In this case, technical terms mean the specific phrases we choose from datasets or metadata of images, and establish as concepts in the network.

To this purpose we are developing the following indices: 1) Bibliographical (volume, issue, page numbers, location, date) to enable statistical analysis of ad publication frequency in one or several newspapers over the course of one or many years. 2) Contextual Informational (brand, product category, company, agency, retailer's address, registered nationality) allowing users to establish a statistical picture of a commodity, in specific newspapers, geographical locations and decades. 3) Content index (sorting by drawing of male, female, elders, youth, middle age people, infant, human, animal, plant, Chinese, foreigner) meaning ad images are hybrid artifacts, mixing text and cartoons; 4) Theoretical categories (the modern, human, woman) to identify categories used aggressively in ads. Once TT in each and every advertisement have been successfully located and the indices of TT identified, we can reveal the APTT of ads, defined as subsets of technical terms that tend to co-occur in an advertisement frequently. With TDM, association pattern discovery can be converted into a statistical inference problem and solved by statistical means.

Second, we seek concept networks that connect key concepts embedded in ads to sociological theories. The Concept Network (CN) is a graph that can efficiently present domain knowledge and reasoning based on it. Each CN node is a concept corresponding to an entity or a technical term. Thus if concept A appears in the definition of concept B, we add a direct link from A to B and domain topology will eventually reflect the structure of the knowledge system: closely related concepts are direct neighbors or locate in the same neighborhood, while concepts belonging to different disciplines or areas will be far away from each other in the graph. Building CN requires indices of concepts and their descriptions. Traditional dictionaries might be a source and our period shows an efflorescence of dictionary publication. Another source is online knowledge databases, like Wikipedia. However domain knowledge of sociological theories are not represented in any language or in any period anywhere on the World Wide Web. To compensate, we are erecting an ontological knowledge database of sociological theories as these appeared in journals, books, articles and the archived documents, 'Social Thought in Modern China, 1830–1940' (STMC). With an ontological database, we can open our sharing platform to define and describe key concepts and relationships among them. Interdisciplinary by design CNMACMS users are welcome to participate by entering data into the databases to improve our model.

Notes

As for the studied have been done by schoalrs on categories like hygienic/卫生), modern/现代, human/人类, eugenics /优生, and female/女性 in commercial/common ideas, please see Jin Guantao(金观涛), Liu Qingfeng(刘青峰), *Studies in History of Idea: The Building of Basic Political Concepts in Modern China* (观念史研究:中国现代重要政治术语的形成), Falv Press, 2009; This book has investigated the origins and transformations of tern basic concepts of "gonghe" (republicanism), "minzhu" (democracy), "quanli" (rights), "geren" (individual), "geming" (revolution), "kexue" (science) in modern Chinese history by using the data of "Database for the Study of Modern Chinese Thought" (1830–1930). Huang Kewu (黄克武), "從申報醫藥廣告看民初上海的意料文化與社會生活 1912–1926" explored the idea of "disease" in advertisements published on *Shen Bao* during early 19th century. Tani Barlow's published papers and book, *In the Event of Women* (Durham: Duke University Press, 2017) establish a historical parallel connecting advertisement ephemera, social theory and the woman category.

Bibliography

Barlow, T. (2012). Advertising Ephemera and the Angel of History. *Positions: asia critique*, **20**: 111–58.

Bergere, M. C. (1989). *The Golden Age of the Chinese Bourgeois, 1911–1937*. Cambridge: Cambridge University Press.

Deng, K., Geng, Z. and Liu, J. S. (2014). Association pattern discovery via theme dictionary models. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), **76**: 319–47. doi: 10.1111/rssb.12032.

Hayles, N. K. (2012). How We Think: Transforming Power and Digital Technologies. In Berry, D. M., (ed.), *Understanding Digital Humanities*. London: Palgrave Macmillan, pp. 42–66.

Manovich, L. (2002). Metadata, Mon Amour, http://manovich.net/index.php/projects/metadata-mon-amour (accessed 14 March, 2016)

The Great War on the Web: the Making of Citing and Referencing by Amateurs

Valérie Beaudouin
valerie.beaudouin@telecom-paristech.fr
Telecom ParisTech, France

Zeynep Pehlivan
zeynep.pehlivan@gmail.com
Telecom ParisTech, France

Over the last two decades, an important amount of work has been carried out by cultural heritage institutions to make their collections available online. How are these digitized collections discovered, discussed and shared on the Web?[1]

The digitized heritage around the First World War (WW1) is an ideal area for such analysis: many digitized collections, centenary anniversary (2014–2018) and an important activism around the Great War. Family, local and militant history are the main motivations of persons that get involved in the history of WW1 (Offenstadt, 2010).

Who are the users who publish or discuss around WW1 on the web? Are they amateurs, experts, academics? How do they publish, share and comment digitized documents? These questions matter for the development of Digital Humanities, but also to those in charge of managing the collections. We conduct an exploratory analysis to understand how the activity around the war is visible in the digital space.

Our research consists of two steps:

1. Identification and categorization of web sites dedicated to WW1 and network analysis of the links between those sources, in order to draw an overall cartography of Web activity on WW1 and to identify the position of amateurs.

2. Focus on one of the main nodes in the web cartography (a forum dedicated to WW1), to analyze the forms of circulations of digitalized documents.

We also conducted a series of semi-structured interviews with participants.

Corpus and methodology

The web is ephemeral: working on web archives allows us to rely on a stabilized corpus, which guaranties the reproducibility of the research (Brügger, 2013). Bibliothèque Nationale de France (BnF), in charge of archiving the French web, created a specific web archive collection dedicated to the centenary of the WW1, based on web sources chosen by librarians. The process of archiving is regularly repeated. Our research relies on the November 2014's archive (9 698 633 Urls for a total size of 800 Gb).

The first step of our analysis consists of generating an oriented network graph on web archives to study the relations (materialized by hyperlinks) between web sources related to WW1. We can consider a link as a pragmatic activity of citing or referencing sources (documents, web sites etc.), although in our approach we are not able to qualify the exact nature and function of the link (Saemmer, 2015).

To make the large-scale graphs readable, the nodes in the graph correspond to the *seed urls* chosen by librarians and all the data crawled from each url is agregated to it. Librarians, in charge of web archiving, qualify the producers (or authors) of the websites. Depending on this categorization, we will distinguish institutional websites (public and official, red) and personal websites (personal and associative, blue).

Mapping WW1 on the Web

The network graph allows to evaluate the place of "amateurs" websites compared to institutional sources.

Basic characteristics of our network are calculated using Gephi (Bastian et al., 2009). The network consisted of 514 nodes and 3713 edges with an average degree of 7.22. The average network distance between all pairs of nodes (average path length) is 2.78 edges with a diameter (longest distance) of 8 edges. The clustering coefficient (the degree to which they tend to cluster together) is 0.27 and the modularity index is 0.28. Overall, WW1 French network is made of highly connected pages (~2.7 edges per node) and shows small-world scale-free network properties (Humphries et al., 2008; Watts et al., 1998) with high clustering coefficient, short average path length and a degree distribution following a "power-law" (smallworldness index = 15).

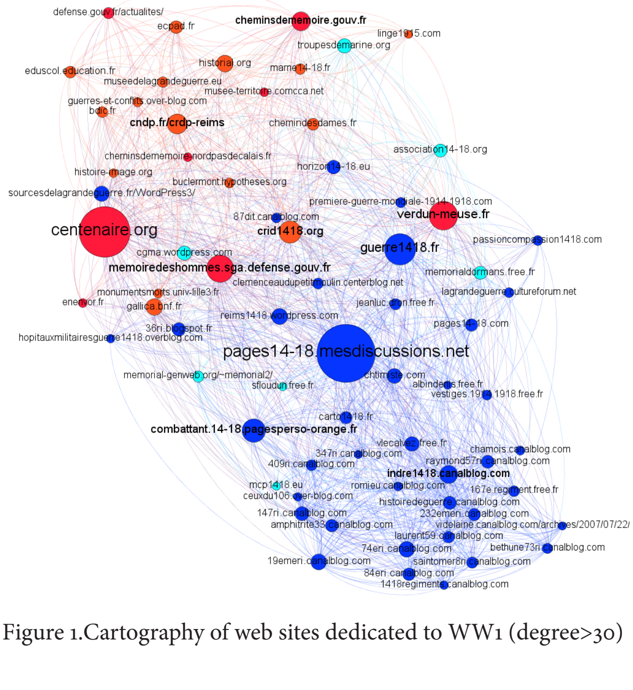


Figure 1.Cartography of web sites dedicated to WW1 (degree>30)

More than half of the sources (52%) come from personal websites that are involved in WW1 as a serious leisure, but not as a profession. The institutional sites are less present (36 %).

To detect influential actors on our network, we use the *degree centrality* which is simply the number of direct relationships that an actor has, the sum of outgoing and incoming links. The network is visualized in Figure 1 by using Gephi, Force Atlas 2 algorithm (Jacomy et al., 2014) as layout with node sizes proportional to their degree centrality. The two main actors are centenaire.org, the official site dedicated to the Centenary, on one side and pages1418.mesdiscussions.net, a web forum managed by amateurs, on the other. Around them, we can distinguish two clusters: the red gathers the institutional sites, while the blue (bottom) gathers all the personal web sites that are intensely interconnected. The forum (Pages 14-18) has a specific position: although it is immersed in the middle of the amateur sphere, it is well connected to institutional sources, because users of this forum are mediators to institutional ressources. The forum constitutes a community of practice (Lave and Wenger, 1991): questions and answers on one side and discussions on the other are two kinds of interactions that allow to share and elaborate knowledge collectively (Conein and Latapy, 2008). Thanks to the forum, a lot of personal websites emerged, each of them dedicated to a specific regiment of foot. Experts of those regiments are the authors of those website that accumate documents (public and personal documents and photographies) on each soldier, each battle, each place occupied.

Degree centrality considers only direct relationships, we also use Hubs and Authorities, known as HITS (Kleinberg, 1999) algorithm. A hub is defined as an actor that points to many other actors and an authority is defined as an actor pointed by many others. HITS algorithm calculates two scores for each actor, hub score and authority score, in a mutually reinforcing way based on the idea that a good authority must be pointed to by several good hubs while a good hub must point to several good authorities (Table 1).

Top 5 Authorities	Authority Score	Top 5 Hubs	Hub Score
memoiredeshommes.sga.defense.gouv.fr	0.0199	pages14-18.mesdiscussions.net	0.0379
centenaire.org	0.0135	centenaire.org	0.0276
crid1418.org	0.0127	guerre1418.fr	0.0251
chtimiste.com	0.0125	combattant.14-18.pagesperso-orange.fr	0.0175
gallica.bnf.fr	0.0124	verdun-meuse.fr	0.0163

Table 1 : Hubs and Authorities

Some authorities, like memoiredeshommes.sga.defense.fr and gallica.bnf.fr, have a very specific profile: as documents warehouses, they receive a lot of links but do not point to other resources. The Centenary website is at the same time a top level authority and hub. The forum, pages1418.mesdiscussions.net, and other personal websites are hubs that point to a relatively large number of authorities.

Forum activity of citations

Based on the specific position of the forum, we decided to focus on it. This forum, founded in 2004 by an amateur, has gradually become a reference site in terms of exchanges and discussions on the WW1. It is a highly active platform with about 400,000 messages in 10 years and about 18,000 subscribers.

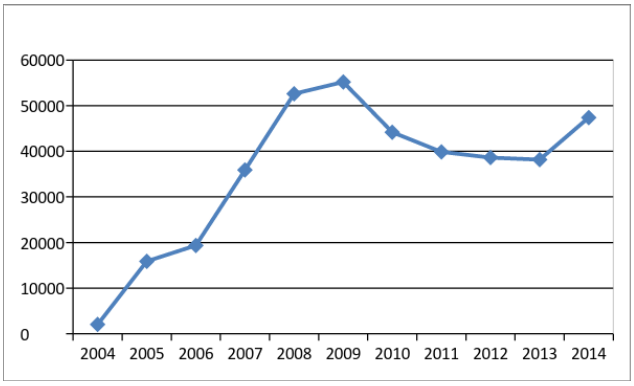


Figure 3. Number of messages published in the forum by year

This forum was archived by BnF in January 2015 and we rely on this corpus for our work. For analysing the activity of citing, we classified the citations into four categories:

- Message_Citation: using a part of previous message
- Quote: text inserted using another source
- Links: hyperlink by using 'link' tag
- Images: hyperlink by using 'img' tag

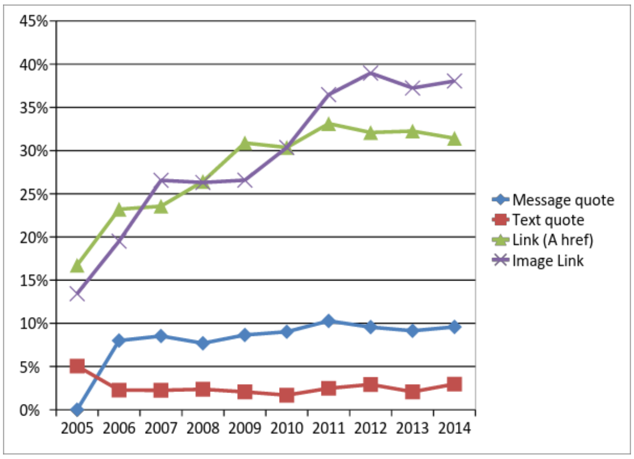


Figure 4. Citations distribution over time

As shown by Figure 4, while the usage of quote and message_citations stays stable, the usage of links and images increases over time according to the increase of digitized documents available on line.

In the corpus, we identified 255,374 image or link citations, an average of 1 citation for 2 messages. We extracted their domain name. The ten more cited domains, which represent 60% of total links, are shown in Table 2.

Netloc	Fréquence
images.mesdiscussions.net	91990
pages14-18.mesdiscussions.net	9636
www.memoiredeshommes.sga.defense.gouv.fr	8984
gallica.bnf.fr	6721
www.asoublies1418.fr	4979
usera.imagecave.com	4274
74eri.canalblog.com	4019
www.servimg.com	2674
imageshack.us	2592
largonnealheure1418.wordpress.com	2407
perso.orange.fr	1916
www.casimages.com	1588
www.memorial-genweb.org	1574
www.hiboox.fr	1492
www.pages14-18.com	1387
pagesperso-orange.fr	1372
perso.wanadoo.fr	1200
albindenis.free.fr	1129
images.imagehotel.net	1041
images4.hiboox.com	1016

Table 2. Most frequent hosts extracted from url citations

The most notable result is the importance of image hosting services. Instead of giving a direct link to an online source, people use hosting services (ex: images.mesdiscussions.net). We can estimate that more that 100 000 citations (40% of total) are of this kind where almost half of them point to images. Users are not confident with the life of web sources. They prefer to download and post the picture than to point to a link which may disappear. The image, in this case is directly available and visible into the post.

Secondly, the forum itself is the most cited website: due to his long history and intense activity, a lot of questions have already been answered. An activity of knowledge management consists in answering a question by signaling an old topic about the same topic.

Memoire des hommes and *Gallica* are the most cited institutional sites. The first one is a gold mine for genealogists who search for ancestors dead during the war and for historians interested in the battles and regiment history. Gallica is a huge warehouse (4 millions of documents): there is a need to identify what kind of documents people are looking for. We extracted the titles of the documents cited and made a content analysis with Iramuteq (Reinert, 1993) (Figure 4); we identified three kinds of documents: photographs and newspapers, official documents ("journaux officiels", "décrets") and documents related to the history of regiments ("historiques des Régiments") from the department of Defense.

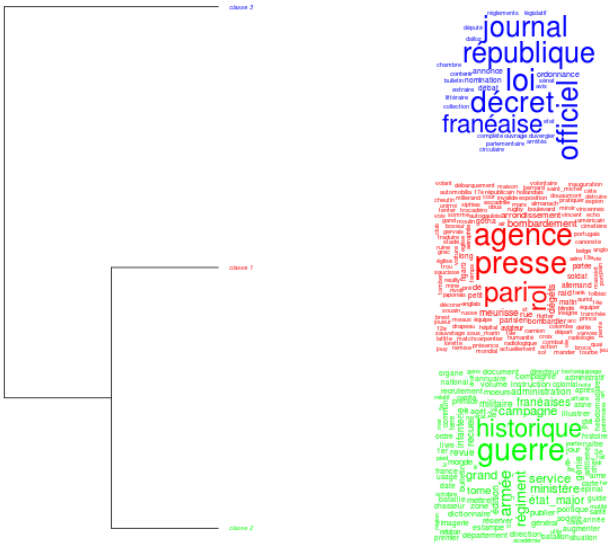


Figure 5. Text Mining clustering of title documents cited from Gallica

Users of the forum also cite a lot of personal web sites dedicated to specific aspects of the war: the history of the soldiers from a specific regiment or squadron for example.

Conclusion

To understand what internet users do with digitized archives, we explored systematically the websites discussing WW1. We developed an original methodology combining web and text mining methods applied to web archives. This method is transferrable. It includes expert advice to qualify the relevant variables for qualifying and selecting sources.

Our analysis shows a great involvment of amateurs (more than half of the websites) in the memory of WW1. They participate to a network of personal websites that gives a specific vision of WW1, more focused on soldiers, regiments, geographic places, objects, remains of war, close to micro-history. A the core of this network, we find the forum, which is the place for interactions and discussions on documents, data, interpretations.

In doing research with digitized resources, amateurs collectively increase their professionalism. Although they do not have the status of academics, they acquire methods in exploring and citing their sources, commenting and sharing with other users. Those amateurs, in doing research with digitized resources, play a major role by discovering and citing institutional heritage documents, adding to their value and creating networks and resources.

Bibliography

Bastian M., Heymann S. and Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks, *International AAAI Conference on Weblogs and Social Media*.

Brügger N. (2013). Historical Network Analysis of the Web, *Social Science Computer Review*, **31**(3): 306-21.

Conein B. and Latapy M. (2008). Les usages épistémiques des réseaux de communication électronique: Le cas de l'Open-Source, *Sociologie du Travail*, **50**(3): 331-52.

Humphries, M.D. and Gurney, K. (2008). Network 'small-world-ness': A quantitative method for determining canonical network equivalence, *PLoS ONE*,**3**(4), e0002051.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* **9**(6), e98679. doi:10.1371/journal.pone.0098679

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness, *Sociometry*, **40**(1): 35-41.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM*, **46**(5): 604-32. DOI=http://dx.doi.org/10.1145/324133.324140

Lave, J. and Wenger, E. (1991). *Situated learning: legitimate peripheral participation*, Cambridge: Cambridge University Press.

Offenstadt, N. (2010). *14-18 aujourd'hui - La Grande Guerre dans la France contemporain*, Paris: Odile Jacob.

Reinert, M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et société*, **66**: 5-39.

Saemmer, A. (2015). *Rhétorique du texte numérique: figures de la lecture, anticipations de pratiques*, ENSSIB.

Watts, D. J. et Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks, *Nature*, **393**(6684): 440-42.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*, Cambridge, U.K.: Cambridge University Press.

Notes

[1] "The future of online digitized heritage: the example of the Great War" is a research project conducted by the BnF, the BDIC and Télécom ParisTech as part of the Cluster of Excellence, Pasts in the present, Investissements d'avenir, réf. ANR-11-LABX-0026-01 (Valérie Beaudouin, Philippe Chevallier, Lionel Maurel, Josselin Morvan, Zeynep Pehlivan, Peter Stirling).

Countering Counter Mapping Methods: Constructing A Humanities GIS Methodology In the Age of Electracy

Clayton John Benjamin
claytonbenjamin@knights.ucf.edu
University of Central Florida, United States of America

Most government and business sponsored geographic information systems (GIS) are employed to manage, track, and exploit labor and production. GIS has primarily been used to regulate political economies, "the concern for life is to administer it (life) through controls and regulations so that resources might be rightly apportioned" (Crampton, 2011). Therefore, governments and businesses use GIS is to produce abstract intelligence and knowledge about people and terrains, which they use to make decisions for distributing goods. These goods can include hospitals, parks, roads, military air strikes, etc. By understanding GIS as an engine for political decision making, we can begin to understand that maps are inherently political. However, GIS, because of its ubiquitous, malleable, and slippery nature, can be used for more than simply calculating statistics to influence the distribution of goods. Instead, GIS can be used to distribute ideas, concepts, emotions, and narratives (the qualitative data of our human actions) in order to act as a counter-map.

Nancy Lee Peluso (1995) first defined counter-mapping as an effort to "appropriate the state's techniques and manner of representation to bolster the legitimacy of "customary" claims to resources" (384). In this definition of counter-mapping, counter-maps contest governmental decisions on political economies by representing location specific knowledge that challenges the abstract statistical knowledge sanctioned by the state. Researchers usually create these counter maps by using participatory GIS (PGIS) methods. PGIS is a cartographic research method that asks members of a population of a mapped geographic area to become active producers in their own mapping. It asks the population members to map their own problems, wants, and needs. For example, Sarah Elwood (2006) has used PGIS with community members from low income neighborhoods in Chicago to help community members advocate for neighborhood improvements (197-208). Though this use of counter-mapping may be helpful to advocate for the redistribution of goods, I argue, it doesn't do enough to contest the primary political functionality of GIS. GIS in these projects often simply replicates the dominant ideology of GIS - that its sole function is to be used to make decisions for the distributions of goods within political economies. This definition of counter-mapping has constrained our imaginations and the ability to play with the functions of GIS and to miss out on potential GIS poetics. The central question in consternation is: how can maps function in the age of electracy unconstrained by the ideologies of the paper map?

Instead of focusing on the political and economic functionality of GIS, which I argue is a print centric way of thinking about GIS, what would it mean to create an electronic poetic discourse of GIS? How can GIS be used to map human emotion, desire, and truth? This type of reasoning resists the temptation to define GIS as a tool for deliberate decision making and to understand GIS as a medium for invention. Invention is the counter logic of GIS hermeneutics -- GIS heuretics. Gregory Ulmer (2002) states, "The purpose of the course is to approach electracy by trying to invent it (what I call "heuretics" --the use of theory to invent forms and practices, as distinct from "hermeneutics," which uses theory to interpret existing works)" (4). How can we invent geography for the electronic apparatus? This question can be answered following the models provided by Ulmer. Ulmer invents a method for applying humanities discourse which relies "not on positivism but quantum relativity; not realism but surrealism." Surrealism, then can act as theory that counters traditional positivistic and "rational" decision making. In the Surrealist Manifesto, Andre Breton (1924) writes:

"(I)n this day and age logical methods are applicable only to solving problems of secondary interest. The absolute rationalism that is still in vogue allows us to consider only facts relating directly to our experience. Logical ends, on the contrary, escape us. It is pointless to add that experience itself has found itself increasingly circumscribed. It paces back and forth in a cage from which it is more and more difficult to make it emerge. It too leans for support on what is most immediately expedient, and it is protected by the sentinels of common sense. Under the pretense of civilization and progress, we have managed to banish from the mind everything that may rightly or wrongly be termed superstition, or fancy; forbidden is any kind of search for truth which is not in conformance with accepted practices" (2).

We can see that, just as I'm searching for a theory and method that resists hermeneutic rationalizations of GIS, so were the surrealists resisting positivist rationalism, because these rationalisms limited creativity and human experience. Therefore, surrealism allows humanities researchers to infuse creativity back into the research milieu.

Furthermore, one humanities method that may be applied to electronic map making, GIS, is psychogeography, an avant-garde method developed by the Situationists who were heavily influenced by surrealism. Guy Debord (1959) defines psychogeography as derive, meaning "to drift." He defines derive as "the study of the precise laws and specific effects of the geographical environment, consciously organized or not, on the emotions and behavior of individuals" (3). By implementing psychogeography, the researcher ignores the boundaries and zones of a city, disturbs the modern distance between researcher and space, and allows the researcher to gain the feeling of a place. By performing psychogeography, the researcher is allowed an intimacy with space and place which may change the conclusions he or she may make about a particular space had he or she relied solely on traditional GIS hermeneutics. Here, emotion has been injected into the research paradigm.

This paper then, describes a method of using GIS tracking services to record a pscyhogeography performed in Bradenton, FL. Currently, there is a heroin epidemic happening in Bradenton, FL, particularly in the South Bradenton neighborhood, and there have been several maps issued by the government and local media agencies that track heroin overdoses in Bradenton, FL. To counter map these official and popular maps, I used Debord's drifting methodology which he outlines in "Theory of the Derive," to perform three different drifts. My goal was to record the feeling of South Bradenton. These drifts were tracked using the GIS interface My Tracks mobile phone application by Google, Inc. During the drifts, My Tracks recorded my path as I drifted through South Bradenton, FL. Additionally, I used my cell phone to take images and videos of the neighborhood and I recorded notes, thoughts, and feelings into a notebook with time stamps. Once the drifts were completed, the My Tracks paths were exported to Google My Maps and the photos and notes were imported into the My Map and added to their corresponding places on the My Tracks path. This presentation will showcase these maps and contrast them to the government maps created about the heroin epidemic. My presentation then focuses on my analysis of the results, the usefulness of psychogeography for geographic research in the digital age, and further research needed for the development of a theory for electronic geographic research and methodologies.

Bibliography

Breton, A. (1924). First Manifesto of Surrealism 1924. In Harrison, C. and Wood P. (eds.) *Art In Theory, 1900-2000 an Anthology of Changing Ideas*. New York: Blackwell Publishing.

Crampton, J. W. (2011).*Mapping: A Critical Introduction to Cartography and GIS*. John Wiley and Sons. vol. **11**.

Debord, G. (1959). Theory of the Derive. In Knabb, K. (ed.) *Situationist international anthology*. Berkeley, CA: Bureau of Public Secrets, pp. 49.

Elwood, S. (2006). Negotiating Knowledge Production: The Everyday Inclusions, Exclusions, and Contradictions of Participatory GIS Research*. *The Professional Geographer*, **58**(2): 197-208.

Peluso, N. L., (1995). Whose Woods Are These? Counter-mapping Forest Territories in Kalimantan, Indonesia. *Antipode*, **27**(4): 383-406.

Ulmer, G. L. (2002). *Internet Invention From Literacy to Electracy*.

Does Character Speech Matter? A Quantitative Approach

Peggy Bockwinkel
peggy.bockwinkel@ilw.uni-stuttgart.de
University of Stuttgart, Germany

Introduction

Character speech is an elementary part of novels. When calculating with German-speaking novels, the question arises, if there is a stylistic difference between character and narrator speech. Given the presupposition that character speech imitates authentic communication, I will answer these questions with a structured series of experiments on deictic expressions, based on statistic and linguistic knowledge.

Theoretical background

Character speech is defined here as direct or cited/ quoted speech, the words and sentences one finds between quotation marks. By drawing on these distinct punctuation markers, it can be separated automatically from the rest of the text. With the beginning of modernism, however, these formal structures have started to dissolve. Between the clear marking of character speech and no marking at all, gradual stages are possible. In these cases, a considerable effort is required to automatically separate character speech from the rest of the novel. An examination of the differences between character and narrator speech therefore can be useful to assess the necessity of such a costly separation.

Deictics are necessary in communication situations to refer to a point in time, space or to certain objects or persons like the speaker or the addressee. In the sentence

Tomorrow I will be there.

every word is deictical. Deictics are context-dependent, which means that in different situations they have different meanings. They belong to different lexical categories – but most of them are function words. The presented approach implies the following premises:

1. Different (non-fictional) texts types show a different frequency of deictic terms. These basic text types have been categorised according to three criteria of communication (dialogical, *face-to-face* and *oral* by Diewald, 1991).
2. Character speech imitates a *dialogue* – which is one of these basic text types.

Transferring the first premise on fiction, literary genre-categorisation has to stand back: character speech in novels and plays would belong to one basic text type.

Previous Work

Previous research on character speech in the humanities shows, that it seems to be a subject mainly in the philosophy of language (see Pafel and Dirscherl, 2015). In literary studies, only one monograph on direct speech exists (Müller, 1981). Recently Brunner (2015) showed that the automated tagging of different kinds of speech in German-speaking novels with computerlinguistic methods is possible, but still has weak results – except for direct speech. Her corpus consisted of 13 different novels, which is a rather small corpus. In computational stylistics it appears that most approaches tend to use smaller corpora (e.g Burrows, 1987; van Dalen-Oskam, 2014).

Method

For the corpus, German-speaking plays and novels from the 18th to the 20th century are selected randomly. They are separated into several subcorpora consisting of 25 plays and 75 novels, including the Brunner corpus with 13 novels (Fig. 1).

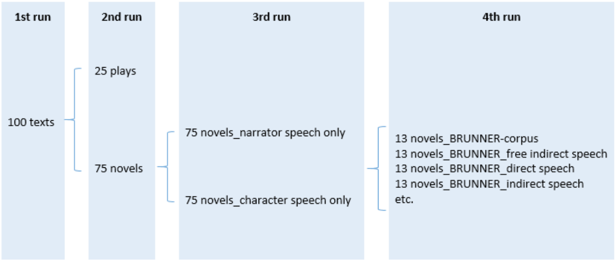

1st run
2nd run
3rd run
4th run
100 texts
25 plays
75 novels
75 novels_narrator speech only
75 novels_character speech only
13 novels_BRUNNER-corpus
13 novels_BRUNNER_free indirect speech
13 novels_BRUNNER_direct speech
13 novels_BRUNNER_indirect speech
etc.

Figure 1: corpus scheme

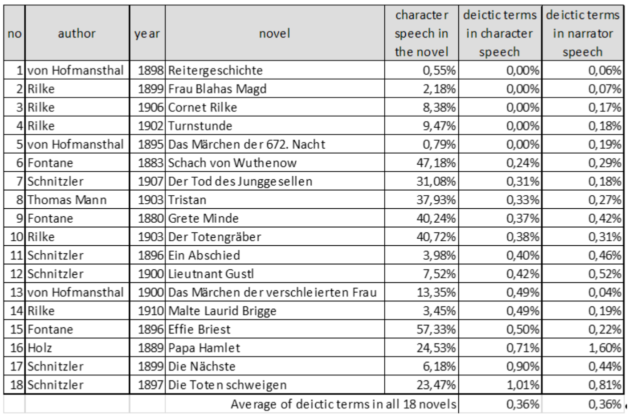
The subcorpus with the plays serves as a benchmark/ reference value: Since plays consist mostly of character speech, they can be compared with the *plays* and *novels_character speech only*-subcorpus. Gries 2008 serves for the statistical basis. Eder's structured experiments in "Does size matter?" (2013) serve as a template for the presented analyses. That means parameters are changed in a consistent and transparent way through the experiment:

The experiment – counting the deictic terms – is run several times (Fig. 1): At first, the whole corpus is tested. For the second run the novels are separated from the plays and for the third run character speech is separated from the rest of the text. Finally, the Brunner corpus is run a forth time with the other categories of speech, e.g. free indirect speech, separated as well.

Since there is no common definition of deictics, nothing like a deictic lexicon exists. In my analyses, I will use a rather straightforward approach: An expression will be regarded as deictic, if it belongs to one of the main deictic categories like time, space, or person and if the deictic reference is its main function. This definition excludes verbs, but includes all function words like temporal and spatial adverbs and a small group of personal pronouns.

Results sample corpus

As a starting point for upcoming research I will present temporary results that are generated by drawing on a small sample corpus of novels in which character speech is separated from the rest of the text and the frequency of deictic terms is analysed (Fig. 2). Then the results are compared with Diewald's results for the basic text types in non-fiction (Fig. 5).



no	author	year	novel	character speech in the novel	deictic terms in character speech	deictic terms in narrator speech
1	von Hofmansthal	1898	Reitergeschichte	0,55%	0,00%	0,06%
2	Rilke	1899	Frau Blahas Magd	2,18%	0,00%	0,07%
3	Rilke	1906	Cornet Rilke	8,38%	0,00%	0,17%
4	Rilke	1902	Turnstunde	9,47%	0,00%	0,18%
5	von Hofmansthal	1895	Das Märchen der 672. Nacht	0,79%	0,00%	0,19%
6	Fontane	1883	Schach von Wuthenow	47,18%	0,24%	0,29%
7	Schnitzler	1907	Der Tod des Junggesellen	31,08%	0,31%	0,18%
8	Thomas Mann	1903	Tristan	37,93%	0,33%	0,27%
9	Fontane	1880	Grete Minde	40,24%	0,37%	0,42%
10	Rilke	1903	Der Totengräber	40,72%	0,38%	0,31%
11	Schnitzler	1896	Ein Abschied	3,98%	0,40%	0,46%
12	Schnitzler	1900	Lieutnant Gustl	7,52%	0,42%	0,52%
13	von Hofmansthal	1900	Das Märchen der verschleierten Frau	13,35%	0,49%	0,04%
14	Rilke	1910	Malte Laurid Brigge	3,45%	0,49%	0,19%
15	Fontane	1896	Effie Briest	57,33%	0,50%	0,22%
16	Holz	1889	Papa Hamlet	24,53%	0,71%	1,60%
17	Schnitzler	1899	Die Nächste	6,18%	0,90%	0,44%
18	Schnitzler	1897	Die Toten schweigen	23,47%	1,01%	0,81%
				Average of deictic terms in all 18 novels	0,36%	0,36%

Figure 2: Percentage of deictic terms (*here*, *now*) in character speech and in the rest of the novel (narrator speech), relative frequency.

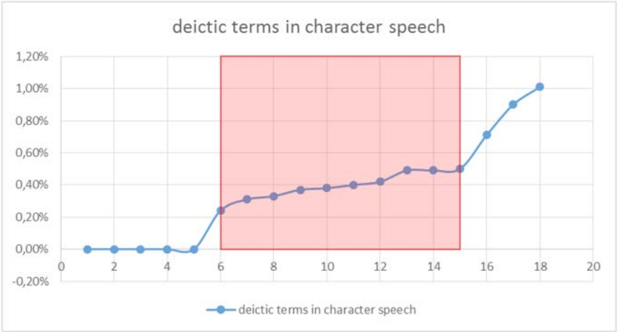


Figure 3: Frequency of deictic terms in character speech of the 18 novels of the table in fig. 2, in ascending order

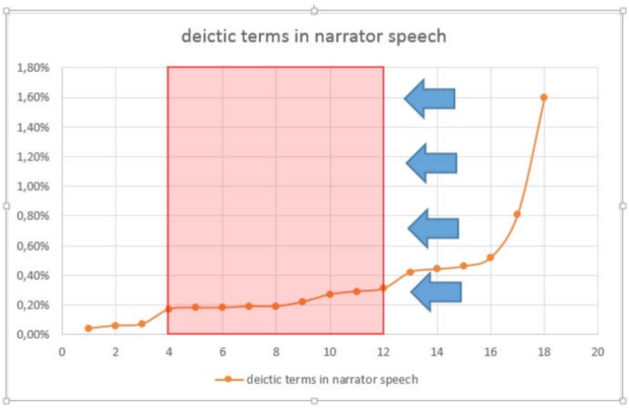
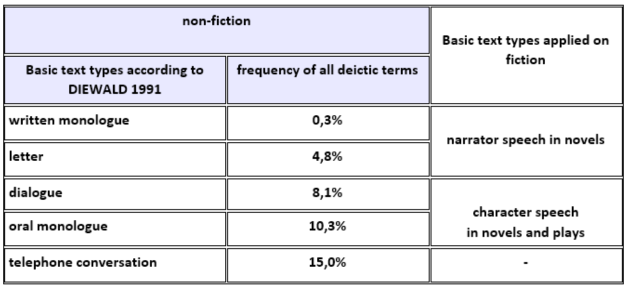


Figure 4: Frequency of deictic terms in narrator speech of the 18 novels of the table in fig. 2, in ascending order

The average values for character and narrator speech (0.36% in either case) show no difference at all (Fig. 2). However, a closer look on the frequency distribution draws another picture: Figure 2 shows that most of the novels (novels 6 to 14) range around 0,4% deictic terms in character speech. In figure 3 the results seen in figure 2 are visualised: in contrast half of the deictic terms in the narrator speech of the novels range a bit higher than 0,2%. This shift is marked by the blue arrows (Fig. 3 + 4). The stability of these results will be evaluated on the complete corpus (see "Methods").

Conclusions

According to Diewald's results the deictic frequency of basic text types is lowest in *letters* and *written monologues*. *Dialogue* and *oral monologue* present a higher frequency of deictic terms, but still topped by *telephone conversation* (Fig. 5). Regarding their principle structure, character speech can be assigned to the basic text types, *dialogue* and *oral monologue*, whereas narrator speech resembles *written monologue* and *letters*.



non-fiction		Basic text types applied on fiction
Basic text types according to DIEWALD 1991	frequency of all deictic terms	
written monologue	0,3%	narrator speech in novels
letter	4,8%	
dialogue	8,1%	character speech in novels and plays
oral monologue	10,3%	
telephone conversation	15,0%	-

Figure 5: Different text types and their frequency of personal, local, temporal and objectual deictic terms, the results refer to Diewald 1991: 383.

Accordingly, the results of the sample corpus indicate a lower deictic frequency in narrator speech compared to character speech, which matches DIEWALD's results on deictic frequency in the basic text types.

Still, for the sample corpus, it is quite problematic to compare these numbers with DIEWALD's results, because in the non-fictional discourses all deictic terms are analysed instead of only two prototypic ones (*here, now*). Since it seems that there is a difference between both types of speech in novels, it nevertheless has to stand the test of the broader experiments with the bigger 100 texts corpus. In addition, for further research another series of this experiment should be run with the same corpus scheme, but set up as a stylometric analysis to see if similar patterns can be found. If the results show a significant difference between character and narrator speech, it is necessary to have character and narrator speech separated in all novels as part of the preparation for digital text analyses.

Bibliography

Brunner, A. (2015). *Automatische Erkennung von Redewiedergabe: Ein Beitrag zur quantitativen Narratologie*, Band 47. Berlin/Boston: de Gruyter.

Bühler, K. (1934/1965). *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Stuttgart: Lucius and Lucius.

Burrows, J. F. (1987). *Computation into criticism: a study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

Diewald, G. M. (1991). *Deixis und Textsorten im Deutschen*. Tübingen: Niemeyer.

Eder, M. (2013). Does size matter? Authorship attribution, small samples, big problem. In *Digital Scholarship in the Humanities* Nov 2013, DOI: 10.1093/llc/fqt066 http://dx.doi.org/10.1093/llc/fqt066.

Gries, S. T. (2008). *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck and Ruprecht.

Müller, A. (1981). *Figurenrede. Grundzüge der Rededarstellung im Roman*. PhD, Göttingen: Georg-August-Universität.

Pafel, J., Dirscherl, F. (2015). *Die vier Arten der Rede- und Gedankendarstellung: Zwischen Zitieren und Referieren. Linguistische Berichte*, **241**: 3–47.

Dalen-Oskam, K. van. (2014). Epistolary voices. The Case of Elisabeth Wolff and Agatha Deken. *LLC: The journal of digital scholarship in the humanities*. **29**(3): 443-51. First published online: May 21, 2014. http://llc.oxfordjournals.org/content/29/3/443.full.pdf?keytype=ref&ijkey=aRCuD3n825cEkmP

Tool-based Identification of Melodic Patterns in MusicXML Documents

Manuel Burghardt
manuel.burghardt@ur.de
Media Informatics Group, University of Regensburg, Germany

Lukas Lamm
lukas.lamm@stud.uni-regensburg.de
Media Informatics Group, University of Regensburg, Germany

David Lechler
david.lechler@stud.uni-regensburg.de
Media Informatics Group, University of Regensburg, Germany

Matthias Schneider
Matthias.Schneider@stud.uni-regensburg.de
Media Informatics Group, University of Regensburg, Germany

Tobias Semmelmann
Tobias.Semmelmann@stud.uni-regensburg.de
Media Informatics Group, University of Regensburg, Germany

Introduction: Digital musicology

Computer-based methods in musicology have been around at least since the 1980s[1]. Besides the creation of digital editions (cf. Kepper et al., 2014; Veit, 2015), scholars in this area of study have also been interested in quantitative approaches for musicological analyses (cf. Müllensiefen and Frieler, 2004; Vigilanti, 2007). Such quantitative analyses rely on music information retrieval (MIR) systems, which can be used to search collections of songs according to different musicological parameters. There are many examples for existing MIR systems, all with specific strengths and weaknesses. Among the main downsides of such systems are:

- **Usability problems**, i.e. tools are cumbersome to use, as they oftentimes only provide a command-line interface and also require some basic programming skills to utilize them; example: Humdrum[2]
- **Restricted scope of querying**, i.e. tools can only be used to search for musical incipits; examples: RISM[3], HymnQuest[4]
- **Restricted song collection**, i.e. tools can only be used for specific collections of music files; various examples of MIR tools for specific collections are described in Typke et al. (2005)

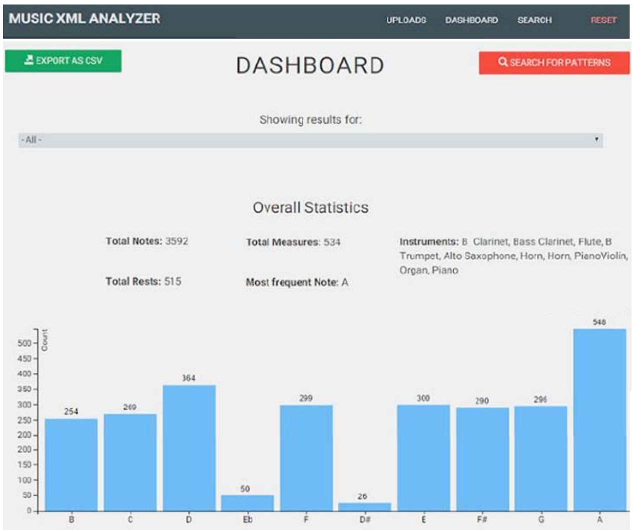
A particularly promising MIR tool can be found in Peachnote[5] (Viro, 2011), which uses optical music recognition (OMR) software to index more than one million sheets from the Petrucci Music Library[6], aiming to provide a search interface for musicology which can be seen as an analog of the Google Books Ngram Viewer[7]. Despite many existing software solutions, we believe that accurate OMR is still a major challenge in digital musicology. At the same time, there are numerous databases[8] at hand, that provide machine-readable music documents, fully annotated with MusicXML (Good, 2001) markup.

On this account, we designed MusicXML Analyzer, a generic MIR system that is trying to overcome the weaknesses of existing MIR tools, and that allows for the analysis of arbitrary documents encoded in MusicXML format.

MusicXML Analyzer: Basic functionality and implementation details

MusicXML Analyzer can be used to analyze songs in a quantitative manner, and to search for specific melodic patterns in a collection of songs. The results of the analyses are rendered as virtual scores and can be viewed in any recent web browser. In addition, the queries and the results can be played as a synthesized audio file; all analyses can also be exported as PDF or CSV files.

The tool comprises three main components: (1) the upload function, (2) the analysis function, and (3) the search function. After one or more files in MusicXML format have been uploaded via an intuitive drag-and-drop dialog, the analysis component parses the data and calculates basic frequencies; the results are stored in an SQL database and can be displayed in a dashboard view (cf. Fig. 1).



Showing results for:

- All -

Overall Statistics

Total Notes: 3592

Total Rests: 515

Total Measures: 534

Most frequent Note: A

Instruments: B Clarinet, Bass Clarinet, Flute, B Trumpet, Alto Saxophone, Horn, Horn, PianoViolin, Organ, Piano

Figure 1: Snippet from the dashboard view, showing basic frequencies for a corpus of MusicXML documents.

The dashboard displays the following information, either for an individual song, or for a corpus of multiple songs:

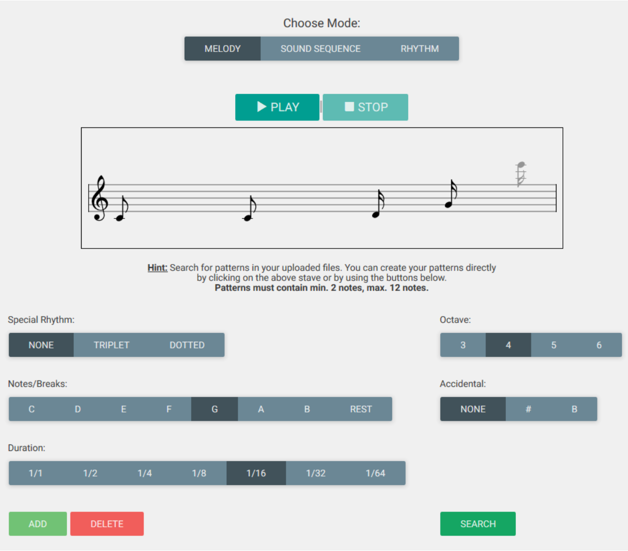
- Overall statistics for single notes, rests and measures
- Types of instruments used in the song (if described in the MusicXML data)
- Frequency distribution for single notes, intervals, keys, note durations and meters

Via a dedicated search function, a corpus of MusicXML documents can be queried for melodic patterns on different levels of information:

- Search for a sound sequence; example: c', c', g', g'
- Search for a rhythmic pattern; example: eighth note, eighth note, quarter note, quarter note
- Search for melodic patterns, i.e. a combination of sound sequence and rhythm; example: c' / eighth note, c' / eighth note, g' / quarter note, g' / quarter note

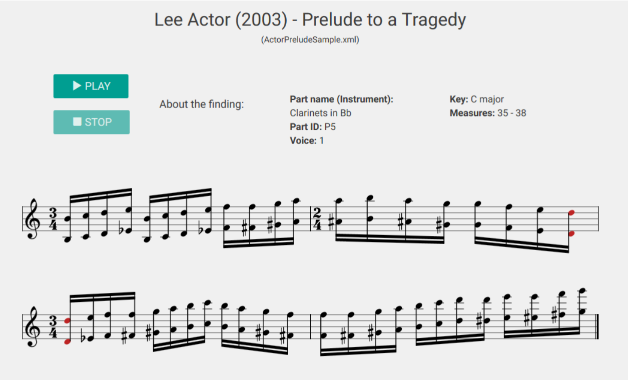
Search queries can be entered via a virtual staff that was realized with the VexFlow library[9] (cf. Fig. 2). Once a search pattern has been entered, it can also be played as a synthesized Midi sequence, which was realized with the Midi.js library[10].

After a query has been submitted, all results – i.e. the songs that contain the search pattern – are displayed in a list view. The list contains the name of the song and also the number of total occurrences of the search pattern in that song. By clicking on one of the song items in the list, a virtual score is rendered for the whole song; the search pattern is highlighted whenever it occurs in that virtual score (cf. Fig. 3). The whole song can be played directly in the web browser, or downloaded for further analyses as a PDF (realized with the jsPDF library[11]).



Hint: Search for patterns in your uploaded files. You can create your patterns directly by clicking on the above stave or by using the buttons below.
Patterns must contain min. 2 notes, max. 12 notes.

Figure 2: Interface for entering queries to identify tonal, rhythmic, or melodic patterns in a corpus of MusicXML documents.



Lee Actor (2003) - Prelude to a Tragedy

(ActorPreludeSample.xml)

About the finding:

Part name (Instrument): Clarinets in Bb
Part ID: P5
Voice: 1

Key: C major
Measures: 35 - 38

Figure 3: Virtual score rendering of a document from the results list; the search pattern is highlighted in red color.

MusicXML Analyzer was implemented by means of standard web technologies (HTML, CSS, JavaScript, PHP), in particular by utilizing the following libraries and frameworks: Laravel[12], jQuery[13], D3.js[14], Bootstrap[15], Typed.js[16], Dropzone.js[17].

A short demo video that showcases the main functionality of the tool is available at https://dl.dropboxusercontent.com/u/4194636/MusicXML-Analyzer.mp4

A fully functional online demo[18] of MusicXML Analyzer is available at http://music-xml-analyzer.herokuapp.com/

MusicXML Analyzer can also be downloaded and modified (according to the MIT open source license) from GitHub: https://github.com/freakimkaefig/Music-XML-Analyzer.

Future directions

In its current implementation, MusicXML Analyzer performs an exact match search, i.e. only documents which have the exact same value in their MusicXML markup will be found by the search function. We are planning to implement a more sophisticated melodic similarity algorithm (cf. Grachten et al., 2002; Miura and Shioya, 2003) that allows for the configuration of different similarity thresholds.

At the same time, we are adapting MusicXML Analyzer for a recent project on a large corpus of German folksongs. Besides monophonic melodies, this collection of folksongs also contains machine-readable metadata (region, date, etc.) and lyrics. Accordingly, we are trying to enhance the search features of MusicXML Analyzer in a way it can not only search songs for melodic patterns, but also for metadata parameters and keywords from the lyrics. Such an enhanced MIR system could be used to analyze the following research questions:

- Are there characteristic melodic and linguistic patterns for German folksongs, from a diachronic perspective as well as from a regional perspective?
- Are there melodic-linguistic collocations, i.e. do certain melodic patterns co-occur with certain keywords or phrases?

Bibliography

Good, M. (2001). MusicXML for Notation and Analysis. In Hewlett, W. B. and Selfridge-Field, E. (eds.), *The Virtual Score: Representation, Retrieval, Restoration*. Cambridge (MA) and London (UK): MIT Press, pp. 113–24.

Grachten, M. A., Josep, L. and Mántaras R. L. (2002). A comparison of different approaches to melodic similarity. *Proceedings of the 2nd International Conference in Music and Artificial Intelligence (ICMAI) 2002*.

Kepper, J., Schreiter, S. and Veit, J. (2014). ‚Freischütz' analog oder digital – Editionsformen im Spannungsfeld von Wissenschaft und Praxis. *Editio*, **28**: 127–50.

Miura, T. and Shioya, I. (2003). Similarity among melodies for music information retrieval. *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM) 2003*.

Müllensiefen, D. and Frieler, K. (2004). Optimizing Measures Of Melodic Similarity For The Exploration Of A Large Folk Song Database. *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR) 2004*, pp. 274–80.

Typke, R., Wiering, F. and Veltkamp, R. C. (2005). A survey of music information retrieval systems. *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR) 2005*, pp. 153–60.

Veit, J. (2015). Music notation beyond paper. On developing digital humanities tools for music editing. *Forschungsforum Paderborn*, **18**: 40-48.

Viglianti, R. (2007). MusicXML: An XML Based Approach to Musicological Analysis. *Digital Humanities 2007: Conference Abstracts*, pp. 235–37.

Viro, V. (2011). Peachnote: Music Score Search and Analysis Platform. *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR) 2011*, pp. 359-62.

Notes

[1] The popular series „Computing in Musicology" started around 1985. For an overview of all volumes of the series cf. http://www.ccarh.org/publications/books/cm/; Note: All URLs mentioned in this text were last checked on March 3, 2016.
[2] http://www.humdrum.org/
[3] https://opac.rism.info/
[4] http://hymnquest.com/
[5] http://www.peachnote.com/
[6] http://imslp.org/
[7] https://books.google.com/ngrams
[8] http://www.musicxml.com/music-in-musicxml/
[9] http://www.vexflow.com/
[10] http://mudcu.be/midi-js
[11] https://parall.ax/products/jspdf
[12] http://laravel.com/
[13] https://jquery.com/
[14] http://d3js.org/
[15] http://getbootstrap.com/
[16] http://www.mattboldt.com/demos/typed-js/
[17] http://www.dropzonejs.com/
[18] Due to some technical limitations of our server environment, the initial access to the online demo may take a few seconds to wake up the server from *idle mode*.

The Preparation of the Topic Model

Rachel Sagner Buurma
rbuurma1@swarthmore.edu
Swarthmore College, United States of America

Replacing "How something is made, with a view to finding out what it is" with "How something is made, with a view to making it again" – the Essence with the Preparation – is linked to an option that's completely antiscientific: in reality, the starting point of the Fantasy [of the critic's writing of a novel] isn't the Novel (in general, as a genre), but one or two novels out of thousands.

-Roland Barthes, *The Preparation of the Novel,* Session of December 9, 1978, 13.

The Literariness of Topic Modeling

This short paper reports on the progress of my attempt to construct a reading of topic modeling using state-of-the-art literary criticism. I argue that dominant digital humanities understandings of topic models assume some of the characteristics of literature most essential to twentieth-century criticism – counter-factuality, a mediated form that is ultimately separable from aesthetic characteristics, and an efficient, self-enclosed, total form. More specifically, I show that topic models also tend to be read by digital humanists according to the assumptions, protocols, and caveats we accord to the interpretation of realist fiction. While often revealing and productive, many digital humanists' uses of topic modeling are indebted to assumptions about the literariness and fictionality of topic models that we have yet to fully understand. Drawing on work by Stephen Ramsay, Johanna Drucker, Alan Liu, and others that theorizes continuities between the values of literary criticism and computational processes, I suggest that we temporarily set aside the idea that topic modeling reveals the "contents" of a set of novels (or of any other corpus). Instead, drawing on Roland Barthes' late work on *The Preparation of the Novel*, we might rethink topics as preparatory notes written by no one, as an imaginary archive whose contents furnish a productively alienating, too-perfect map of the novel's preparation. In *Preparation*, Barthes moved away from his earlier work's emphasis on totalizing interpretations of literature's meaning to think about models of the text that allow for a more partial and slow view of the process of meaning creation. Topic modeling has the potential for helping us towards a Barthesian reimagination of the novel's reader as the novel's writer, of the search for the fantasy origins of a novel as a method that pulls us away from formal totality and a form-content divide. While this reorientation comes out of literary studies, I also suggest that it might have applications for more instrumental uses of topic modeling outside the realm of the humanities, in which assumptions about topics as equivalent to a document set's "contents" also tend to draw on our conventions for reading realist genres.

Fictionality and the Topic Model

The past few years have seen the rapid popularization of topic modeling among humanist scholars in general, and among scholars of literature in particular. The literature on topic modeling abounds in stern and salutary warnings about the limits and dangers of topic modeling for humanistic study. One can read about the dangers of introducing algorithmic black boxes into literary research, the concern that literary scholars are unprepared to fully (or even partially) interpret the topic models and their related data, and the worry that they fail to understand even the interpretive choices made during corpus preparation. Part of the worry derives from a larger assumption that topic modeling "reveals" the "contents" of novels. We assume that literary critics dipping their toes into topic modeling will shed their traditional interpretive caution in the face of the algorithm's authority, and will misunderstand the un-semantic nature of topics or accept meaningless correlations as meaningful. I want to suggest that all such warnings are relevant only given a very limited understanding of what a topic model is, its imagined relation to the corpus from which it derives, and the goals of the model's interpreter. These warnings do usefully help us think about some of the invisible interpretive choices we make when we choose chunk and clean documents, apply stoplists, select a number of topics to train, and - most importantly, assign semantic labels to unsemantically generated topics. And yet these warnings assume either that topic models aspire to be mimetic maps of the corpuses they model or that technologically unsophisticated interpreters of topic models imagine that this is the case. This is not surprising; the assumption that topic models are a realist genre is pervasive in literature on topic modeling, literary and otherwise. Yet if we relieve ourselves of this constraint and instead substitute a more plausible frame – the topic model's fictionality – we will be able to enjoy a wider range of relations between model and corpus.

In place of assuming that topic models belong to the realm of realism, then, we might pay more attention to the generative uncertainty of topic modeling and to its literal fictionality. Topics are probabilistically-created formations, and the algorithm that generates topic models is based on the enabling – but crucially, counterfactual – "assumption that documents have multiple topics." (Boyd-Graber et al., 2015). By looking at the documents we offer it, the algorithm generates topics that, in given proportions, compose each document. (Or, rather, it generates the probability that a certain percentage of words in every given document were generated by a given particular topic.) Topics, of course, don't actually exist prior to the documents that generate them; they don't actually exist independently in the same way the documents at all. They are, in a certain sense, fictions. Topics are things that might have existed – but didn't! - given the existence of the document set in question. While we can and sometimes do relegate this fact to the realm of methodology, the fictionality of topics is crucial to remember for any literary-critical uses of topic modeling, for it reminds us that these models offer us a view of our document set radically at odds with any other more literal sources of a novel we might use – such as an author's notes towards a novel, or a catalog of the virtual or actual library of books a novelist brings to the writing table, or even the looser sense of social "discourses" that exist prior to novels and which we might imagine in part "composing" a novel. Using a few targeted examples drawn from topic models of corpuses of nineteenth-century novels of varying sizes and comparing them to some examples of nineteenth-century novelists' notebooks, I suggest that reimagining topic models as fictional notes might be not just a theoretical exercise but a practical way of conceptualizing the relation between topic model and corpus.

Bibliography

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, **55**(4): 77. doi:10.1145/2133806.2133826.

Blei, D. M. (2014). Topic Modeling and Digital Humanities. *Journal of Digital Humanities* (April 8, 2013). http://journalofdigitalhumanities.org/2-1/topicmodeling-and-digital-humanities-by-david-m-blei/

Belvins, C. (2010) Topic Modeling Martha Ballard's Diary. April 1, 2010. http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/

Boyd-Graber, J., Mimno, D. and Newman, D. (2014). Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In Airoldi, E. M., Blei, D., Erosheva, E. A. and Fienberg, S. E. (eds), *The Handbook of Mixed Membership Models and Their Applications*. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2015.

Buurma, R. S. and Heffernan, L. (2014). Notation After 'The Reality Effect': Remaking Reference with Roland Barthes and Sheila Heti. *Representations*, **125**(1): 80–102. doi:10.1525/rep.2014.125.1.80.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, **22**: 288–96.

Erlin, M. (2014). The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731-1864, In Erlin, M. and Tatlock L. (eds.), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Rochester, NY: Camden House, pp. 55-90. .

Goldstone, A. and Underwood, T. (2014). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History: a journal of theory and interpretation*, **45**(3).

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Jockers, M. L. and Mimno, D. (2013). Significant Themes in 19th-Century Literature. *Poetics*, **41**(6): 750–69. doi:10.1016/j.poetic.2013.08.005

Laudun, J. and Goodwin, J. (2013). Computing Folklore Studies: Mapping over a Century of Scholarly Production through Topics. *Journal of American Folklore*, **126**(502): 455-75.

Meeks, E. (2013). The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities*, **2**(1). http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/

Ramsay, S. (2011). *Reading Machines: Towards an Algorithmic Criticism*. Urbana, Chicago, Springfield: University of Illinois Press.

Rhody, L. M. (2013). Topic Modeling and Figurative Language. *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Rhody, L. M. (2016). The Story of Stopwords: Topic Modeling an Ekphrastic Tradition. *Digital Humanities 2014*, Lausanne, Switzerland (accessed January 3, 2016). http://www.lisarhody.com/the-story-of-stopwords/

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities*, **2**(1). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

Tangherlini, T. R. and Leonard, P. (2013). Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research. *Poetics*, **416**: 725-49. doi:10.1016/j.poetic.2013.08.002.

Notes

[1] See Buurma, R.S. and Heffernan, L. (2014). "Notation After 'The Reality Effect': Remaking Reference with Roland Barthes and Sheila Heti." *Representations* 125:1 80–102. doi:10.1525/rep.2014.125.1.80.

[2] Blei, D. M. (2014). Topic Modeling and Digital Humanities. *Journal of Digital Humanities* (April 8, 2013). http://journalofdigitalhumanities.org/2-1/topicmodeling-and-digital-humanities-by-david-m-blei/ Erlin M (2014). The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731-1864, p 55-90. In Matt Erlin (ed. and introd.) and Lynne Tatlock (ed. and introd.), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Rochester, NY: Camden House; Goldstone, A. and Underwood, T (2014). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History: a journal of theory and interpretation* 45:3; Jockers, M.L. and Mimno, D (2013). Significant Themes in 19th-Century Literature. *Poetics* 41:6, 750–69. doi:10.1016/j.poetic.2013.08.005; Laudun, J. and Goodwin, J. (2013). Computing Folklore Studies: Mapping over a Century of Scholarly Production through Topics. *Journal of American Folklore* 126:502, 455-475; Meeks, E (2013). The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities*, 2:1. http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/; Rhody, L.M. (2013). Topic Modeling and Figurative Language. *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/; Rhody, L.M. (2016). The Story of Stopwords: Topic Modeling an Ekphrastic Tradition. *Digital Humanities 2014,* Lausanne, Switzerland. Accessed January 3, 2016. http://www.lisarhody.com/the-story-of-stopwords/ and Tangherlini, T.R. and Leonard, P (2013). Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research. *Poetics* 41:6 (December 2013): 725-749. doi:10.1016/j.poetic.2013.08.002.

[3] Benjamin Schmidt warns, for example, that "simplifying topic models for humanists who will not (and should not) study the underlying algorithms creates an enormous potential for groundless — or even misleading — "insights."" Schmidt worries that a pair of assumptions about topic models – that they are "coherent" and "stable" – "let humanists assume that the co-occurrence patterns described by topics are meaningful; topics are useful because they describe things that resemble "concepts," "discourses," or "fields."" Schmidt is worried, that is, that the appearance of semantic meaning we find in "good" topics will seduce humanists into thinking that they have discovered the "contents" of novels – whereas what topic modeling really offers us is exactly a non-semantic machine indexing of a set of texts about which our approaches tend to be based on ground assumptions about semantic meaning. Benjamin Schmidt, "Words Alone: Dismantling Topic Modeling in the Humanities." *The Journal of Digital Humanities* 2:1 (Winter 2013), http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

[4] One recent paper, for example, describes good topics with the example of "trout fish fly fishing water angler stream rod flies salmon…" explaining that the topic "is specific. There is a clear focus on words related to the sport of trout fishing. It is coherent. All of the words are likely to appear near one another

in a document. Some words (water, fly) are ambiguous and may occur in other contexts, but they are appropriate for this context. It is concrete. We can picture the angler with his rod catching trout in the stream. It is informative. Someone unfamiliar with the topic can work from general words (fishing) to learn about more unfamiliar words (angler). Relationships between entities can be inferred (trout and salmon both live in streams and can be caught in similar ways)." (Boyd-Graber J, Mimno D and Newman D (2014) Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In: Airoldi E M, Blei D, Erosheva E A, Fienberg S E (eds.), *The Handbook of Mixed Membership Models and Their Applications*. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2015.)

[5] As Boyd-Graber et alia note, "Topic models are based on a generative model that clearly does not match thway humans write. However topic models are often able to learn meaningful and sensible models." (2014: 15).

Éditions Critiques Électroniques Et Structuration Du Contenu Sur Les Plateformes De Consultation Numériques: Normes Et Pratiques

Joana Casenave
joana.casenave@gmail.com
Université de Montréal Canada, Université Paris-Est Créteil France

Yves Marcoux
yves.marcoux@umontreal.ca
Université de Montréal Canada, Université Paris-Est Créteil France

Introduction et cadre conceptuel

L'édition critique électronique, part active des humanités numériques, est aujourd'hui en pleine expansion, comme en témoigne la multiplication de projets d'envergure dans le domaine[1]. Parallèlement à la création de ces projets, les éditeurs électroniques effectuent également un travail important de définition de ce champ disciplinaire. La réflexion épistémologique, développée par les théoriciens, est également nourrie par les praticiens. Multiforme, le travail de définition d'un champ disciplinaire en cours de structuration et d'évolution est primordial pour asseoir la discipline sur des bases conceptuelles solides. Ainsi, dans chacune des monographies récentes consacrées aux humanités numériques, la question de l'édition critique électronique est traitée et débattue[2].

Pour définir l'édition critique électronique, les chercheurs disposent du terreau historiographique et conceptuel façonné par l'édition critique traditionnelle et la philologie. L'édition critique électronique se situe en effet au lieu de rencontre de la philologie et des techniques informatiques. L'évolution apportée par les nouvelles technologies dans ce domaine est significative. Les humanités numériques, auxquelles se rattache l'édition électronique, impliquent de nouveaux modes de recherche, transdisciplinaires et collaboratifs, dans lesquels les techniques informatiques ont toute leur part (Burdick et al., 2012). Les éditeurs électroniques, justement, considèrent que l'imprimé n'est plus le vecteur de transmission principal du savoir. Selon eux, le web peut contribuer à améliorer la diffusion des connaissances. Ainsi, ces éditeurs tirent parti des outils technologiques afin d'améliorer et d'approfondir l'accès au texte pour le lecteur. (Vanhoutte, 2010).

Normes de structuration des éditions critiques traditionnelles et leurs évolutions numériques

Les éditions critiques traditionnelles, imprimées sur papier et regroupées dans des collections spécialisées[3], répondent à des codes stricts de structuration du contenu. C'est un jeu complexe de normes de lisibilité, élaborées au fil des siècles, qui a donné au texte imprimé son maximum d'efficacité et en a permis une lecture facile et rapide (Vandendorpe, 1999). Ces normes ont été mises en place au fil du temps, en fonction des pratiques de lecture et des règles établies par les éditeurs. Elles sont de divers ordres et portent sur différents aspects du texte : disposition et découpage du texte dans la page, orthographe, ponctuation ou encore syntaxe. L'homogénéité que l'on perçoit dans les éditions critiques est due à l'édiction de règles très précises qui permettent de guider le chercheur dans chacune des étapes de l'établissement du texte. L'apparat critique, cœur du travail du philologue, comporte lui aussi ses codes de structuration particuliers (Duval, 2006).

Lorsque l'on passe au support numérique, ces conventions disparaissent en partie. A l'écran, il s'agit donc de reprendre les normes élaborées pour l'imprimé et de les faire évoluer afin qu'elles s'adaptent aux particularités du format numérique (Souchier, 2012). Les éditeurs qui s'occupent des livres et des artefacts scientifiques numériques, dont font partie les éditions critiques électroniques, doivent donc d'une part reprendre certains codes sémiotiques du livre imprimé pour les adapter au format numérique, et d'autre part inventer de nouveaux codes propres à l'environnement numérique.

Problématique

Il existe un standard dévolu à l'encodage des documents textuels dont font partie les éditions critiques : il s'agit bien entendu du langage XML/TEI. Mais si la structuration des données est régie par les règles de la TEI, la présentation finale de l'information n'a pas fait, pour sa part, jusqu'à ce jour, l'objet d'une quelconque normalisation. Dans notre communication, nous nous attacherons précisément à cette question de structuration et de présentation du contenu sur les plateformes de consultation des éditions électroniques. Ce qui nous intéresse est de savoir quelles sont les règles empiriques de structuration, de disposition et de présentation de l'information qui peuvent être observées dans les éditions électroniques. Comment est-ce que les normes de présentation du texte édité, de l'apparat et des outils de recherche ont évolué de l'imprimé à l'écran ? Nous ferons une étude sémiotique de l'édition critique électronique. Ainsi, nous pourrons voir si l'édition critique électronique actuelle est, ou non, en voie d'inventer de nouveaux codes de structuration de l'information et de lecture du contenu.

Méthodologie et résultats attendus

Pour étudier cette question, nous analysons un corpus d'éditions critiques électroniques sélectionnées par choix raisonné. Nous avons constitué un corpus de douze éditions qui illustrent la diversité des démarches éditoriales observées dans le champ des éditions électroniques : éditions critiques originellement publiées sur papier et transposées en format numérique, éditions critiques publiées simultanément de manière électronique et papier et enfin éditions critiques électroniques natives. Elles ont été développées par divers types d'institutions : universités (Ecole nationale des chartes ou Université de Virginie par exemple), instituts de recherche (à l'image de l'Institut Huygens) ou maisons d'édition (Sd-editions notamment). Enfin, certaines de ces éditions électroniques sont le fait d'initiatives personnelles de chercheurs, comme c'est le cas pour l'édition électronique des œuvres de Petrus Plaoul. Pour mener notre étude, nous nous fondons sur l'observation de ces éditions électroniques ainsi que sur les recherches déjà menées dans ce domaine, incluant les études les plus récentes.

Les éditions critiques électroniques retenues pour notre corpus d'observation sont diverses : elles couvrent les traditions anglo-saxonnes et françaises et portent sur des documents d'archives ou des œuvres littéraires. La variété des types d'éditions critiques que nous avons retenues répond d'abord à un souhait d'examen le plus complet possible des points centraux de l'édition critique. Nous allons donc présenter un état des lieux des éditions critiques électroniques, au travers d'une analyse de la structuration du contenu.

Pour mener notre étude, nous avons élaboré une grille d'observation. Cette dernière est concentrée sur trois pôles : l'apparat critique, le paratexte et les outils de navigation et recherche proposés aux lecteurs. C'est en fonction du positionnement initial de l'éditeur électronique que nous allons comparer les normes de structuration et de présentation de l'information des éditions critiques électroniques. En effet, le positionnement éditorial joue un rôle important dans la fonction assignée à l'édition électronique. Dès lors, des tendances normatives s'observent au sujet de la structuration du contenu dans chacune des catégories d'éditions électroniques (éditions natives ; éditions traditionnelles au départ et publiées ensuite de manière électronique ; éditions simultanément publiées de manière traditionnelle et électronique). Bien entendu, le public de l'édition critique participe aussi, à tout le moins par ses attentes, à l'élaboration des éditions électroniques. Ainsi, la place du public et le rôle qu'il tient sont abordés à chacun des points d'observation (apparat critique, paratexte, outils).

Les observations que nous ferons nous permettront de dégager des tendances et caractéristiques des éditions critiques électroniques afin de comprendre comment ces éditions sont organisées et comment s'y créent des normes de structuration de l'information.

Bibliography

Apollon, D., Belisle, C. and Régnier, P. (2014). *Digital Critical Editions*. University of Illinois Press.

Burdick, A., et al. (2012). *Digital humanities*. Cambridge: MIT Press.

Duval, F. (dir.) (2006). *Pratiques philologiques en Europe: actes de la journée d'étude organisée à l'École des chartes le 23 septembre 2005*. Paris: École nationale des chartes.

Gold, M (dir.) (2012). *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Pierazzo, E. (2015). *Digital Scholarly Editing* . Ashgate.

Sahle, P. (2015). Répertoire des éditions critiques électroniques. http://www.digitale-edition.de/

Souchier, E. (2012). La « lettrure » à l'écran. Lire et écrire au regard des médias informatisés, *Communication et langages*, **174**: 85-108.

Terras, M., Nyhan, J. and Vanhoutte, E. (2013). *Defining Digital Humanities. A Reader.* Farnham: Ashgate Publishing.

Vandendorpe, C. (1999). *Du papyrus à l'hypertexte. Essai sur les mutations du texte et de la lecture*. Montréal: Boréal et Paris, La Découverte.

Vanhoutte, E. (2010). Defining electronic editions: a historical and functional perspective. *Text and Genre in Reconstruction. Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: Open Book Publishers, pp. 119-44.

Warwick, C., Terras, M., Nyhan J. (2012). *Digital humanities in practice*. London: Facet Publishing in association with UCL Centre for Digital Humanities.

Notes

[1] Un répertoire des éditions critiques électroniques est tenu par Patrick Sahle ; il compte plus de 360 projets qui varient aussi bien par le type de sources éditées que les langues et époques concernées (http://www.digitale-edition.de/).

[2] Citons par exemple *Digital humanities* (Burdick et al., 2012), *Debates in the Digital Humanities* (Gold, Klein, 2012), *Digital humanities in practice* (Warwick et al., 2012), *Defining digital humanities* (Terras et al., 2013) qui consacrent tous des pages ou des chapitres à l'édition critique électronique. En outre, certaines publications récentes sont plus particulièrement dédiées à l'analyse des éditions critiques électroniques elles-

mêmes : *Digital Critical editions* (Apollon et al., 2014) et *Digital Scholarly Editing* (Pierazzo, 2015).

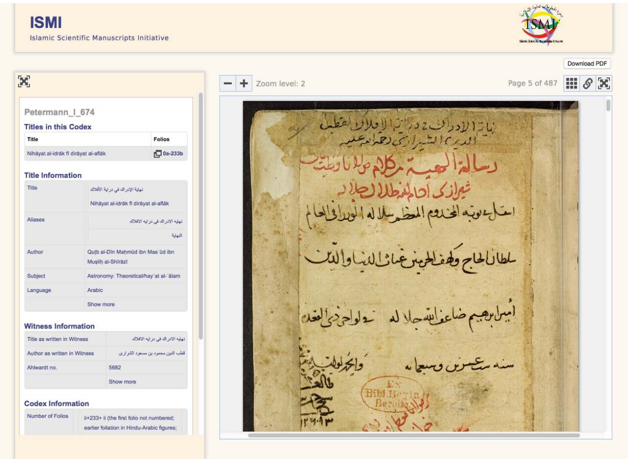
[3] Prenons l'exemple de la maison Honoré Champion, qui propose les trois collections suivantes dédiées aux éditions critiques: *Classiques français du Moyen âge*, *Textes littéraires de la Renaissance*, *Textes de littérature moderne et contemporaine*.

An Islamic Manuscript Database as a Network of Objects

Robert Casties
casties@mpiwg-berlin.mpg.de
MPIWG, Germany

The ISMI project

The Islamic Scientific Manuscripts Initiative (ISMI) is a project by the Max Planck Institute for the History of Science and the Institute of Islamic Studies at McGill University in Montreal to collect and make accessible information on all Islamic manuscripts in the exact sciences (astronomy, mathematics, optics, mathematical geography, music, mechanics, and related disciplines), whether in Arabic, Persian, Turkish, or other languages ranging in time from the 8th to the 19th century. Since 2007 the project has collected information on over 4000 texts existing in 14500 witnesses in 7500 codices, as well as information on over 2000 persons. A preliminary website with a subset of information on 130 codices is already available online (https://ismi.mpiwg-berlin.mpg.de).



Petermann_I_674

Titles in this Codex

Title	Folios
Nihāyat al-idrāk fī dirāyat al-aflāk	0a-233b

Title Information

Title	نهاية الإدراك في دراية الأفلاك Nihāyat al-idrāk fī dirāyat al-aflāk
Aliases	نهايه الادراك فی درايه الافلاك النهاية
Author	Quṭb al-Dīn Maḥmūd ibn Mas'ūd ibn Muṣliḥ al-Shīrāzī
Subject	Astronomy: Theoretical/hay'at al-'ālam
Language	Arabic

Witness Information

Title as written in Witness	نهايه الادراك فی درايه الافلاك
Author as written in Witness	قطب الدين محمود بن مسعود الشرازى
Ahlwardt no.	5682

Codex Information

Number of Folios	ii+233+ ii (the first folio not numbered; earlier foliation in Hindu-Arabic figures;

Figure 1: Online view of codex Peterman I 674 (Staatsbibliothek Berlin)

The project's goal is to create a catalogue database of all relevant manuscripts and record as much information on these manuscripts as available. The collected data contains basic bibliographic information but also paleographic and codicological information, and also information about the content of the texts and information about the uses of the manuscript and its users.

The manuscripts sometimes and notes and colophons providing information about the reading of a text or the use of a text in teaching, about sponsors and the acquisition and ownership of the manuscripts over time. With this information the database not only provides a powerful bibliographical research tool for scholars in the field, but also helps to answer questions pertaining to the historical and social context of knowledge like: Was the author working as an isolated individual or as part of a scientific group? Was this a well-known text? Did it influence subsequent workers in the field? Was it studied at a court or in a school?

A manuscript database

Cataloging old manuscripts is already a task that overwhelms standard bibliographical databases. There is the problem of anonymous authors at the same time as a proliferation of authors with the same or similar names. The integration of information from different sources is made difficult by libraries changing the names of their collections and their numbering systems or libraries themselves being incorporated or centralized into other institutions. Adding to that are the problems of handling of Arabic writing and the multitude of slightly different Arabic romanization systems.

Other requirements also arose in the project early on: for example the need to record and present outdated information. In many cases authorship information has been misattributed widely and for long times so that scholars coming to the database looking for specific information may search under a wrong name or assume the database to be in error unless the common misattribution is also presented with arguments of why the old information is superseded.

Manuscripts as network of objects

The database development started in 2006 with a new data model based on the idea of a network of flexible objects and relations. Objects can have arbitrary attributes which are text strings. The relations between objects are also like objects and also have attributes.

The objects are things like an abstract *TEXT*, a concrete *WITNESS* and a real or imaginary *PERSON* while relations like *is_exemplar_of* connect a text and its witnesses and *was_created_by* connects the texts and a person as its author. The same person can at the same time also be connected to other witnesses as a copyist or as a sponsor.

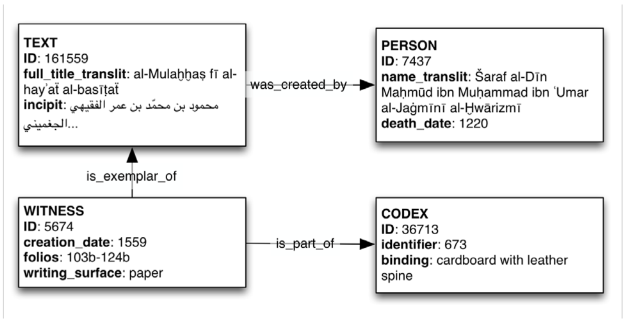

TEXT
ID: 161559
full_title_translit: al-Mulaḫḫaṣ fī al-hay'aẗ al-basīṭaẗ
incipit: محمود بن محمّد بن عمر الفقيهي الجغميني...
was_created_by
PERSON
ID: 7437
name_translit: Šaraf al-Dīn Maḥmūd ibn Muḥammad ibn ʿUmar al-Jaġmīnī al-Ḫwārizmī
death_date: 1220
is_exemplar_of
WITNESS
ID: 5674
creation_date: 1559
folios: 103b-124b
writing_surface: paper
is_part_of
CODEX
ID: 36713
identifier: 673
binding: cardboard with leather spine

Figure 2: Data model showing relations between text, witness, person and codex objects.

The flexible nature of the relations made it easy to introduce new relation types as it became necessary in the research process, for example to record the documented reading of a manuscript or the misattribution of authorship.

This concept of a network of objects with flexible relations, also called an attribute-graph exists in database products like Neo4J today but those were not available in 2006 which led to the development of a custom database called "OpenMind". The database software is Open Source, written in Java, uses a conventional SQL database backend and a Web-based frontend.

Challenges of networked data

The network-like structure of data in the database makes it easy to add new relation types or new attributes to objects while at the same time making it more difficult to create simple forms for entering data for things that are composed of multiple objects in the data model. A form for a manuscript for example not only creates a witness object but also creates relations to a text object, multiple person objects, codex, collection, library and place objects, creating those objects if they do not exist.

Visualising and querying such networks of data objects is also a new challenge where few established tools and concepts exist. Data can be projected into conventional tools like tables and spreadsheets researchers may be familiar with but these do not exploit the full potential of existing relations. Network visualisation tools and methods on the other hand make if often difficult to browse and search for specific items and require a careful selection of semantically relevant relations for the application of standard graph-theoretical measures.

The project currently explores different tools and methods and gathered input from expert scholars in the field in a workshop in February 2016 to be presented at the conference.

Bibliography

Ragep, Jamil F., and Sally P. Ragep. (2008). The Islamic Scientific Manuscript Initiative (ISMI) Towards a Sociology of the Exact Sciences in Islam. In Calvo E., Comes M., Puig R., and Rius M. (eds.), *A Shared Legacy: Islamic Science East and West. Homage to Professor J. M. Millàs Vallicrosa*, Barcelona: University of Barcelona, pp. 15–21. https://www.rasi.mcgill.ca/ISMI_SharedLegacy.pdf

Deux Projets D'Édition Numérique Dans Le Cadre Du Projet SyMoGIH: Le Journal De Léonard Michon Et Les Actes Des Synodes Des Églises Réformées De Bourgogne

Christine Chadier
christine.chadier@univ-lyon3.fr
Université Jean Moulin - Lyon 3, France

Rosemonde Letricot
rosemondeletricot@hotmail.fr
Université Jean Moulin - Lyon 3, France

Francesco Beretta
francesco.beretta@ish-lyon.cnrs.fr
CNRS UMR 5190 LARHRA

Sylvain Boschetto
sylvain.boschetto@ish-lyon.cnrs.fr
CNRS UMR 5190 LARHRA

Le portail d'édition numérique de sources, http://xml-portal.symogih.org/, que nous présentons aux DH2016 au travers de deux projets d'édition numérique n'est pas un outil de visualisation d'images ou de textes comme pourraient l'être des systèmes de type Omeka ou Drupal. Il s'agit de mettre à disposition de manière dynamique à la fois le texte d'une source mais aussi les données d'une base relationnelle qui constituent son apparat critique. Il faut concevoir ce portail d'édition davantage comme une brique de développement du projet symogih.org.

Le projet symogih.org (Système Modulaire de Gestion de l'Information Historique) a développé un modèle générique de stockage des données historiques permettant leur interopérabilité et leur publication [http://symogih.org/, tous les sites web ont été consultés le 30 octobre 2015]. À partir de ce modèle a été mis en place un système d'information collaboratif pour la recherche en histoire qui est aujourd'hui utilisé par 15 projets de recherche et environ 50 utilisateurs individuels. La plateforme du projet symogih.org offre un outillage numérique accessible et pérenne pour le stockage et la publication de données extraites de l'étude de documents archivistiques et bibliographiques.

Il est possible d'intégrer des données de nature variée qui décrivent l'activité humaine, sociale, économique ou intellectuelle rassemblant, autour d'événements datés et sourcés, des acteurs individuels ou collectifs, des concepts ou des objets géographiques. Le système autorise également l'articulation de ces données avec des textes codés en XML – traités selon les recommandations de la Text Encoding Initiative [https://groupes.renater.fr/wiki/symogih/symogih_manuel/edition_de_textes_en_xml-tei] – ou encore la mise en relation avec des images et leurs métadonnées. La réalisation d'un système d'information géographique (SIG) [http://geo-larhra.org/] joue un rôle essentiel dans ce modèle afin d'associer à ces différents objets leur "empreinte spatiale" et ainsi permettre des analyses spatiales diachroniques.

L'un des derniers développements du projet symogih.org a été la mise en ligne d'un portail de publication des éditions numériques élaborées au sein du LARHRA. Si symogih.org permettait aux chercheurs de mettre à disposition les données structurées issues de leur travail analytique, le portail d'édition rend désormais possible la contextualisation d'une source écrite grâce à ces mêmes données. Les textes encodés en XML/TEI sont stockés sur un serveur eXist-db à partir duquel sont développés différents services (visualisations du texte et des données, géolocalisation, possibilité d'export, moteur de recherche). Le point de jonction entre le texte numérique et la base de données se situe à l'intérieur même des balises TEI où sont intégrés les identifiants du système d'information qui font le lien avec les données structurées. Des fonctionnalités de visualisation peuvent être développées à partir des textes encodés stockés dans une base de données native XML, offrant davantage d'interactivité avec les données et augmentant l'expérience de lecture. De plus, grâce aux technologies web, via les langages XQuery + HTML/Javascript, le portail d'édition numérique permet de collecter des données au-delà du référentiel commun de symogih.org directement à partir du web de données (DBpedia, IdRef, ...).

À ce jour, la partie publique du portail recueille deux projets, l'un d'édition de sources (les *Mémoires* de Léonard Michon), l'autre d'annotation sémantique et de contextualisation de documents concernant l'histoire des savoirs scientifiques à l'époque moderne (Society religion science) [http://xml-portal.symogih.org/web-publications.html]. Un troisième projet d'édition de documents est en cours d'élaboration et verra bientôt le jour : l'édition des Actes des synodes des églises réformées de Bourgogne.

L'édition numérique des *Mémoires* de Léonard Michon fait partie d'une recherche doctorale [Letricot, Rosemonde. *Édition critique numérique des* Mémoires *de Léonard Michon (1715-1746)*. Sous la direction de Hours, Bernard. Université Jean Moulin Lyon 3, LARHRA UMR5190] qui vise à mettre à disposition du public et de la communauté scientifique l'édition critique du Journal historique d'un notable lyonnais relatant la vie des élites bourgeoises de la ville de Lyon de la première moitié du XVIIIe siècle. Le travail d'encodage XML/TEI s'est principalement centré sur l'identification des segments d'information et des entités nommées (personnes, institutions, lieux, etc.) ce qui permettra de recourir à des analyses quantitatives sur les pratiques d'écriture (fréquence, récurrence de noms, etc.) et sur la nature des informations données dans l'ouvrage, que nous pourrons ensuite traduire en parcours de lecture pour le public (parcours thématique, biographique, chronologique, etc.).

Quant aux Actes des synodes des églises réformées, ils représentent une source essentielle pour la connaissance du protestantisme français sous l'Ancien Régime. Ces assemblées réunissent régulièrement des représentants de toutes les églises d'une province pour traiter des affaires qui leur sont communes : questions financières, disciplinaires, doctrinales, etc. Si les sources sur le protestantisme français font l'objet d'une édition chez Droz dans une sous-série de la collection "Travaux d'Humanisme et Renaissance" intitulée "Archives des Églises réformées de France", il n'y avait pas de projet en Humanités numériques sur le sujet [Une édition "papier" des Actes des Synodes Provinciaux des Églises Réformées est en cours chez Droz, le premier volume édité par Didier Boisson a été publié en 2012 et concerne l'Anjou-Touraine-Maine (1594-1683). Le second volume proposera les actes des églises de Bourgogne et sera édité par Yves Krumenacker]. Ce sera bientôt le cas avec l'édition sur le portail XML du projet symogih.org des Actes des synodes des églises réformées de Bourgogne au XVIIe siècle, réalisée sous la direction de Yves Krumenacker de l'Université Jean Moulin Lyon 3.

Ces deux projets d'édition ne se limitent pas à la seule publication de sources mais proposent une édition enrichie par des renseignements récoltés dans des sources complémentaires, saisis dans la base collaborative du projet symogih.org et utilisés non seulement pour préciser l'un ou l'autre renseignement fourni par le texte, selon la démarche d'annotation classique d'une édition papier, mais en proposant également une explication contextuelle dynamique – que permet l'édition numérique – en croisant toute sorte de données et tout en conservant leur traçabilité, par l'enregistrement des sources et de la bibliographie pour chaque information. Il est ainsi possible de reconstituer, par exemple, la généalogie et la carrière d'un pasteur, ou sa bibliographie [L'utilisation d'une base de données collaborative et cumulative permettant de multiplier les sources : journal de pasteur (Bernus, Auguste (1888). Le ministre Antoine de Chandieu d'après son journal autographe inédit 1534-1591. *Bulletin historique et littéraire publié par la Société de l'histoire du protestantisme français*, Tome XXXVII), sources régionales (Papillon, Philibert (1742). *Bibliothèque des auteurs de Bourgogne*. Dijon: Philippe Marteret), *Registres de la Compagnie des pasteurs de Genève*, édités chez Droz].

Les lieux mentionnés dans les écrits et dans la documentation annexe, sont renseignés et géolocalisés dans le gazeteer du projet symogih.org. À partir des données spatiales encodées dans le texte XML, on pourra ainsi réaliser des cartes interactives illustrant différents aspects du codage : villes organisatrices des synodes, lieux d'origine des pasteurs, carte des églises absentes, parcours professionnels des pasteurs. Des fonctions interactives permettront de rebondir de la carte vers les textes ou les informations respectives.

Nous soulignerons lors de la présentation l'apport de la visualisation dynamique des documents pour les deux éditions, chacune avec ces spécificités. En particulier, l'intégration directe des données de la recherche permet une plus grande interactivité : l'apparat critique peut être à tout moment complété au fil des découvertes des chercheurs par la mobilisation simultanée de données provenant d'un silo d'information commun. De plus, les interfaces de visualisation spatiale, ou les graphes mettant en évidence les relations entre les textes et leurs contenus, facilitent l'accès aux documents édités. Les nombreuses possibilités d'enrichissement et d'exploitation des textes amènent les chercheurs en histoire, mais aussi le public, à une découverte sous d'autres angles et avec de nouvelles perspectives du contenu des textes édités.

Nous souhaitons montrer l'apport pour la recherche historique d'un portail d'édition numérique de sources, en insistant sur les bénéfices d'une utilisation croisée de textes encodés en XML/TEI et d'une base de données collaborative. Les projets présentés permettent de retracer le processus de traitement de l'information, de la source à son édition numérique dynamique.

Constructing Evidence in the Photographic Archive: The Experience of Digital Humanists

Alexandra M Chassanoff
achass@email.unc.edu
School of Information and Library Science, University of North Carolina Chapel Hill

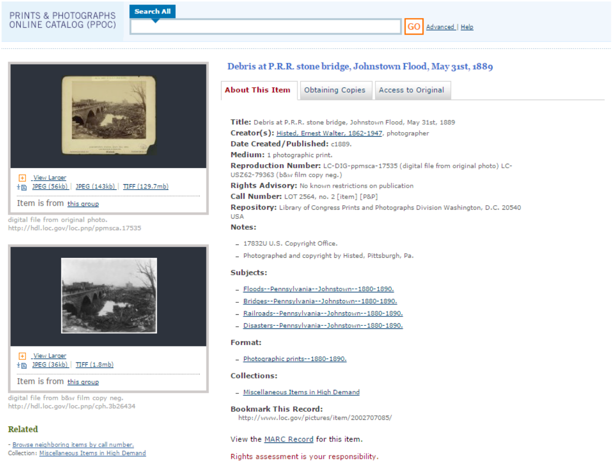
Widespread digitization of cultural heritage materials has presented scholars with unprecedented access to primary sources. For digital humanities scholars, who craft arguments from examination of primary sources, increased access to materials has been celebrated as the "democratization of historical research" (Bolick, 2006). Presumably, such changes in archival research environments have influenced how humanist scholars work. Indeed, recent research has confirmed that technological advancements have significantly impacted scholarly practices (Rutner and Schonfeld, 2012; Chassanoff, 2013). Yet *how* humanists evaluate and use digitalsource materials to construct narratives is less well-understood.

In this paper, I report on findings from a one year qualitative study examining digital humanists' scholarly use of one kind of digital source material - digitized archival photographs. Using a case study approach, I examine the practices and processes at play in the construction of historical evidence. This research is guided by the following questions:

- How and why are digital humanists using digitized archival photographs in their research and teaching activities?
- What factors and qualities matter to them in their experiences in the photographic digital archive?

The goal of this study is to provide an in-depth, holistic understanding of a complex interaction space made up of, but not limited to: digital surrogates of archival objects, user perceptions and attitudes, environmental constraints, and historical training and orientation. Empirical research in digital environments tends to focus on single components of the interaction (e.g., user and interface; user and artifact) as they relate to specific aspects of information behavior, or to conceptualize information use as the successful fulfillment of stated information needs. Yet such perspectives do not attend to the impact that ecological factors may have on user interactions with materials. Adopting a phenomenological stance enables a focus on understanding "how persons construct meaning" through examinations of their particular experiences with certain phenomenon (Wilson, 2002). In this study, exploring *how and why* scholars use digital photographs helps to reveal the emergent qualities and attributes that make this experience meaningful for participants.

Semi-structured interviews were conducted with sixteen participants (9 men and 7 women) throughout the spring and summer of 2015. Recruited participants came from a variety of academic departments, including History, English, African-American Studies, American Studies, Classical Studies, and Musicology. Each participant preselected two examples of digitized archival photographs they had used in research and teaching. Two customized web pages, which I termed *Photograph Scenarios*, were created in an attempt to replicate where possible their original experience viewing and encountering the digital photograph. I also collected supplementary materials related to each participant's image use, including conference presentations, class presentations, course syllabus, dissertation chapters, and journal articles. Data were analyzed using open coding and thematic analysis in order to surface salient aspects of the experience related to interpretation and use. To strengthen and verify the analysis, triangulation across data sources was employed.


PRINTS & PHOTOGRAPHS ONLINE CATALOG (PPOC)
Search All
GO
Advanced | Help
About This Item
Obtaining Copies
Access to Original
View Larger
JPEG (56kb) | JPEG (143kb) | TIFF (129.7mb)
View Larger
JPEG (36kb) | TIFF (1.8mb)

Debris at P.R.R. stone bridge, Johnstown Flood, May 31st, 1889

Title: Debris at P.R.R. stone bridge, Johnstown Flood, May 31st, 1889
Creator(s): Histed, Ernest Walter, 1862-1947, photographer
Date Created/Published: c1889.
Medium: 1 photographic print.
Reproduction Number: LC-DIG-ppmsca-17535 (digital file from original photo) LC-USZ62-79363 (b&w film copy neg.)
Rights Advisory: No known restrictions on publication
Call Number: LOT 2564, no. 2 [item] [P&P]
Repository: Library of Congress Prints and Photographs Division Washington, D.C. 20540 USA

Notes:

- 17832U U.S. Copyright Office.
- Photographed and copyright by Histed, Pittsburgh, Pa.

Subjects:

- Floods--Pennsylvania--Johnstown--1880-1890.
- Bridges--Pennsylvania--Johnstown--1880-1890.
- Railroads--Pennsylvania--Johnstown--1880-1890.
- Disasters--Pennsylvania--Johnstown--1880-1890.

Format:

- Photographic prints--1880-1890.

Collections:

- Miscellaneous Items in High Demand

Bookmark This Record:
http://www.loc.gov/pictures/item/2002707085/

View the MARC Record for this item.

Rights assessment is your responsibility.

Item is from this group
digital file from original photo.
http://hdl.loc.gov/loc.pnp/ppmsca.17535

Item is from this group
digital file from b&w film copy neg.
http://hdl.loc.gov/loc.pnp/cph.3b26434

Related
- Browse neighboring items by call number.
Collection: Miscellaneous Items in High Demand

Figure 1: Photograph Scenario used in interview from Library of Congress collection (http://www.loc.gov/pictures/item/2002707085/)

The findings presented in this paper shed light on both practical concerns and intellectual challenges that surface throughout the experience of constructing historical evidence. Descriptive quotations from interview transcripts alongside examples of image use illustrate the functional ways humanists are using digitized archival photographs in their scholarly activities (e.g., to make historical assertions, to corroborate existing information, etc.). A typology illustrating how scholars use digital photographs is presented below.

Humanist used photograph to...	**Interview excerpt**
Corroborate existing information	"A lot of times we need to look at these historic photos to actually know what it actually looked like and not just our idea of what it looked like"
Make historical assertions	"And I was having trouble finding certain types of ferries that I know existed because of other records. You know, like you would find references to their being a rope ferry, but then you could never find a photo to see exactly what they meant by that."
Reference historical documentation	"Because we were working on just one of the bridges, and as you can see from where they had, it's on the well, the left and the right side, it shows the different bridges that were there."
Elicit reactions from viewer	"But I'd always try to find good touristy photos to show some good 60s and 70s touristy photos or whatever, just to make it livelier"
Present community perceptions at time of creation	"…it's sort of like what did the community, or the boosters of the community, think was important? Because, hey, it's what they were trying to put out for the public…"
Juxtapose against other sources	"Yeah, and that's exactly what I did, is I put it in conversation with other images like it."

Table 1: Typologies of Use

The cases presented demonstrate the extent to which material conditions of *experience* (rather than the tools through which users discovered or accessed resources) can impact interpretive practices and further use.

One salient theme is the factors that motivate participants to use photographs. In one case study, a participant discovered a "vernacular body of images" that led them to collect similar types of visual sources. In another case, a participant selected their project topic after encountering multiple published images depicting what they term "radical masculinity" in the Communist party. They describe their discovery experience: "But then I started noticing there was a lot of- not this image in particular- but there were several other ones that were repeatedly published, of Communists bandaged and showing black eyes and stuff, after confrontations with the police or vigilantes, so that was the first thing I noted."

Other cases illustrate the discrepancies between viewing digital photographs online and encountering them in person. In one case, a participant describes the advantages of obtaining a high quality scan from an archivist rather than accessing and using the "tiny tiny little print" displayed in the online collection. In another case, a participant attributes the discovery of a number of details that became central to their historical argument only after viewing a physical version of the photograph in person:

> So, there are a whole number of things that I never would have seen [if I hadn't also viewed the photograph in person], I think. And a whole range of- you know, the physical image has a whole affect to it that you don't get from a glowing screen.… I can look at the digital image from the [library], and zoom and zoom in to see it, but I never would have seen [this important detail] if I was just looking at a 2 inch by 3 inch digital version.

Both of these examples demonstrate the importance of individual scholars' interpretive experiences as a means for understanding the conditions that enable further use of these primary sources.

Digital humanities scholars undoubtedly face a number of practical and hermeneutic challenges in using source materials from digital archival research environments. As spaces of knowledge production, online archives must support numerous heterogeneous practices in order to remain useful and relevant. At the same time, navigating online archives to find and use sources requires digital humanists to be competent at varying levels, including: interfaces, digitization quality, overall orientation to the archive, and domain-specific heuristics (Yakel and Torres, 2003).

While there has been unquestionable growth in access to digital archival materials, there have been few empirical attempts to understand the factors and qualities that matter to scholars as they interact with digital sources. The case studies presented in this paper will inform design and development efforts in the digital humanities community by focusing on the end-users of these systems and their needs. A holistic, methodological approach which emphasizes scholars' experiences evaluating, interpreting and using materials in scholarly activities allows for an exploration of the mediating factors underlying these practices. Such an orientation enables us to extend our attention beyond simply exploring the material constraints of resource discovery and access, which are often modeled on analog approaches to communication. Instead, this study broadens the emphasis to, in the words of Gregory Bateson, "the difference that makes a difference."

Bibliography

Bateson, G. (1972). *Steps to an Ecology of Mind.* Chicago: University of Chicago Press.

Bolick, C. (2006). Digital archives: Democratizing the doing of history. *The International Journal of Social Education*, **21**(1): 122.

Chassanoff, A. (2013). Historians and the Use of Primary Source Materials in the Digital Age. *The American Archivist*, **76**(2): 458-80.

Collier, J. and Collier, M. (1986). *Visual Anthropology: Photography as a Research Method.* Albuquerque: University of New Mexico Press.

Rutner, J. and Schonfeld, R. C. (2012). Supporting the Changing Research Practices of Historians . *Final Report from ITHAKA S+R Research Publications.*

Stone, S. (1982). Humanities scholars: Information needs and uses. *Journal of Documentation*, **38**(4): 673-91.

Wilson, T. D. (2002). Alfred Schutz, phenomenology and research methodology for information behaviour research. *Proceedings of the Fourth International Conference on Information Seeking in Context.* Lisbon: Universidade Lusiada.

Yakel, E. and Torres, D. (2003). AI: Archival intelligence and user expertise. *The American Archivist*, **66**(1): 51–78.

Compiling a Database on Historical China from Local Records: The Local Gazetteers Project at MPIWG

Shih-Pei Chen
schen@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Zoe Hong
zhong@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Dagmar Schäfer
dschaefer@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Martina Siebert
msiebert@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Jorge Urzúa
jurzua@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

This paper introduces a digital humanities project at the Max Planck Institute for the History of Science (Max-Planck-Institut für Wissenschaftsgeschichte, MPIWG) that aims to unlock the treasure chest of local knowledge written in the genre of Chinese local gazetteers for computer assisted analyses.

In the past two decades, a great amount of historical documents have been digitized and put on the Web to enable easy access for scholars around the globe. In parallel, the amount of searchable full-text versions of historical texts has increased, which opens the possibility of text mining the contents for large scaled analyses. Many works in this direction have been proposed and got recognition, while they are also criticized for drawing conclusions from seemingly imprecise results due to the restriction that their algorithms have no knowledge about the meanings of the different pieces in a text (Jockers, 2013; Google Ngram Viewer; Chen et al., 2007).

An alternative approach is thus to first "teach" computers what each pieces of a text means before asking computers to run automatic analyses. Such "teaching" is done by tagging, or called markup. Many digital humanities projects have been using TEI, a standard for text encoding based on XML, to tag their research materials (TEI; Flanders).

In the Local Gazetteers Project, since the genre organizes knowledge in a very structural way, we are also using tagging to teach computers the meanings of texts in order to turn them into data tables to enable computer assisted analyses including GIS mapping. However, since the amount of texts is huge, we also propose a research data repository for scholars to collaborate in this project and to aggregate their results.

What are the Chinese Local Gazetteers?

The Chinese local gazetteers is a genre of texts that has been produced in China consistently from the 10th century on to even today. Most of them are compiled by local officials as a major means to collect and aggregate historical, social, and geographical knowledge of an administrative region for governing purposes. There are at least 8,000 titles of pre-1949 local gazetteers still extant today. They cover almost every well-populated region in historical China.

Despite being compiled by different officials for different regions, the local gazetteers have developed a pretty consistent structure of "describing" local knowledge. Most gazetteers contain the following chapters: history, geography, local government, infrastructure (buildings, schools, temples, bridges), local products (grains, plants, animals, drugs, commodities), people (local officials and celebrities), and literature. The vast number, the longue durée, the width in geographical coverage, together with the extensive and consistent selection of topics have made the local gazetteers major sources for knowledge about regions for scholars from later periods. To date, several digitization projects on the local gazetteers have been conducted by either commercial vendors or public institutions, and thousands of titles have been made available in searchable full texts for scholarly access.[1]

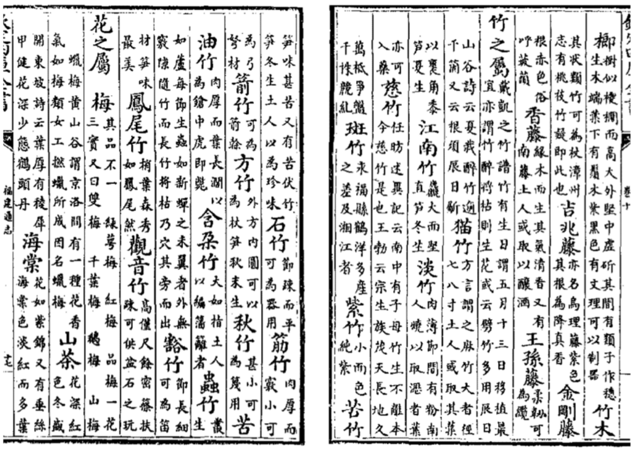
The Project's Goal

Local gazetteers are well studied, but the vast amount of information contained within also makes scholars struggled to study them analytically. We noticed that the local gazetteers often organize knowledge by listing the items: list of temples, lists of flora and fauna, lists of local officials, etc. (See Figure 1 for an example.) This characteristic makes the local gazetteers a database by nature.

The goal of this project is to transform this genre into a scholarly enhanced database in order to enable new forms of digital historical analyses. By turning texts (of lists of things) into data (tables), it will be much easier to aggregate the rich knowledge written in all the extant local gazetteers to compile a database for historical China on local knowledge across geographical regions and time periods. Research-oriented queries, visualization of query results, and further large scaled analyses will then be more easily realized with such a database.

Our approach

We have identified and developed a set of digital tools to help this process. They are: (1) an **extraction interface** that helps historians to tag and to transform digital texts and their built-in structures into *data tables*; (2) a **research data repository** where historians can *share and publish* their data collected via the extraction interface; (3) a set of **digital analytical, visualization, and analysis tools** that can be applied on the collected data for posing research questions.



椰樹似椶櫚而高大外堅中虛斫其間有類子作穗生木端葉下有鬚木紫黑色有文理可以制器竹木
其狀類竹可為杖漳州志有桃枝竹疑即此也古兆藤亦名烏理藤紫色其根為降真香金剛藤
根赤色俗呼菝葜香藤緣木而生其氣清香又有南藤土人或取以釀酒王孫藤柔韌可為纜
竹之屬戴凱之竹譜竹有生日謂五月十三日移植最宜亦謂竹醉將枯則生花或云劈竹多用辰日
山谷詩云要栽醉竹過十筍又云根須辰日斲猫竹方言謂之麻竹大者徑七八寸土人或取其葉
以叢角黍筍夏生江南竹叢大而堅直節冬生淡竹肉薄節間有粉南人燒以取瀝者葉
亦可入藥慈竹任昉述異記云南中有子母竹生不離本令慈竹是也王勃云宗生族茂天長地久
鵝柢爭蟹千株朧亂斑竹永福縣鶴洋多產之差及湘江者紫竹小而色純紫苦竹

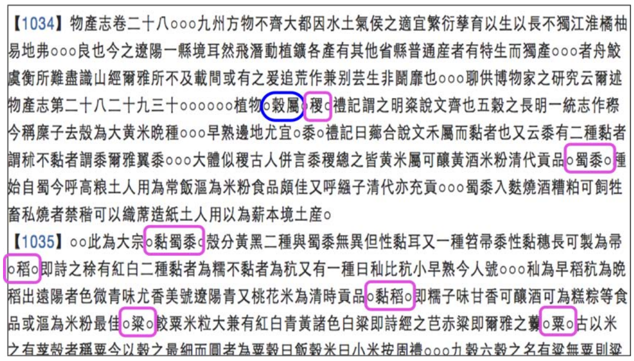
筍味甚苦又有苦伏竹筍冬生土人以為珍味石竹節疎而平可為器用筋竹肉厚而竅小可
為弓弩材箭竹可為箭笴方竹外方內圓可以為杖筍秋末生秋竹甚小可為篾用苦
油竹肉厚而葉長潤以為鎗中虎即斃含朶竹大如指土人以編籬落者蟲竹叢生
如蘆每節生蟲如蟖蟬之未翼者外無竅隨竹而長竹將枯乃穴其旁而出豁竹節長細可為笛
材筍味最美鳳尾竹梢葉森秀如鳳尾然觀音竹高僅尺餘密篠扶疎可供盆石之玩
花之屬
梅其品不一綠萼梅紅梅品梅一花三實又曰雙梅千葉梅鴛梅山梅
蠟梅黃山谷謂京洛間有一種花香氣如梅類女工撚蠟所成因名蠟梅山茶花深紅色冬藏
開東坡詩云葉厚有稜犀甲健花深少態鶴頭丹海棠花如紫綿又有垂絲海棠色淡紅而多葉
福建通志

Figure 1 Two pages from the chapter of "local products" from *Fujian tongzhi* (Qing, 1737). The pages list items according to categories. The bigger fonts represent the names of the items, while the smaller fonts are descriptions of the items. (Image source: http://ctext.org/library.pl?if=en&file=50305&page=80.)

I. An extraction interface for transforming digital texts into data

Due to the well-formatted organization of the local gazetteers, some scholars have tried to write computer programs to parse the digital texts in order to collect data for their own research (Mitchell, 2015; Chang, 2015). However, due to the fact that each of the local gazetteers has slightly different formats, it is difficult to write generic computer programs that can work on all the gazetteers. The China Biographical Database project (CBDB) tackled this problem by building an interactive user interface that allows scholars to describe observable writing patterns of the text in regular expressions (Wikipedia, 2015) – The Smart Regex Tool. The compiled regular expressions can be saved, run against other texts, and modified to fit slightly different writing patterns. CBDB used this interface and efficiently collected 250,000 records on local officials from 290 local gazetteers using only 420 man-hours (Pang et al., 2014).

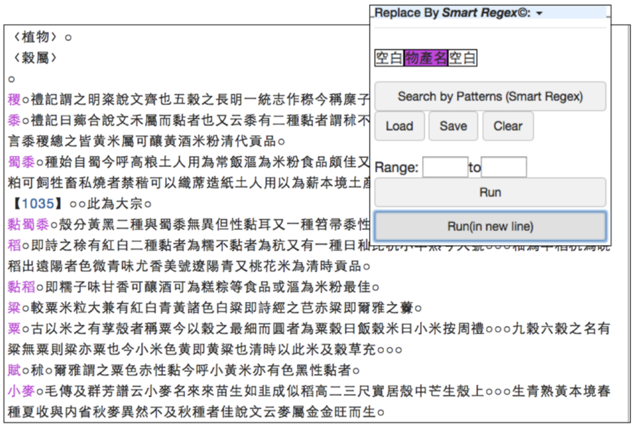
We inherited this interface from CBDB and further improved it so that scholars can define different tag sets and regular expressions to capture information relevant to topics they are interested in. Figure 2 to Figure 6 are the step-by-step screenshots of how we use this interface to tag a text on local product and to export the result into a table. The resulting table contains not only the names, categories, descriptions, alternative names, and usages of the products, but also the source gazetteer, the chapter name, and page number as well as geographical coordinates for mapping purpose.



【1034】物產志卷二十八○○○九州方物不齊大都因水土氣侯之適宜繁衍孳育以生以長不獨江淮橘柚易地弗○○○良也今之遂陽一縣境耳然飛潛動植鑛各產有其他省縣普通産者有特生而獨產○○○者舟鮫虞衡所難盡識山經爾雅所不及載間或有之爰追荒作兼别芸生非闕靡也○○○聊供博物家之研究云爾述物產志第二十八二十九三十○○○○○○植物○穀屬○稷○禮記謂之明粢說文齊也五穀之長明一統志作穄今稱糜子去殼為大黃米晚種○○○早熟邊地尤宜○黍○禮記日薌合說文禾屬而黏者也又云黍有二種黏者謂秫不黏者謂黍爾雅翼黍○○○大體似稷古人併言黍稷總之皆黃米屬可釀黃酒米粉清代貢品○蜀黍○種始自蜀今呼高粮土人用為常飯漚為米粉食品頗佳又呼繦子清代亦充貢○○○蜀黍入麩燒酒糟粕可飼牲畜私燒者禁稭可以織蓆造紙土人用以為薪本境土産○

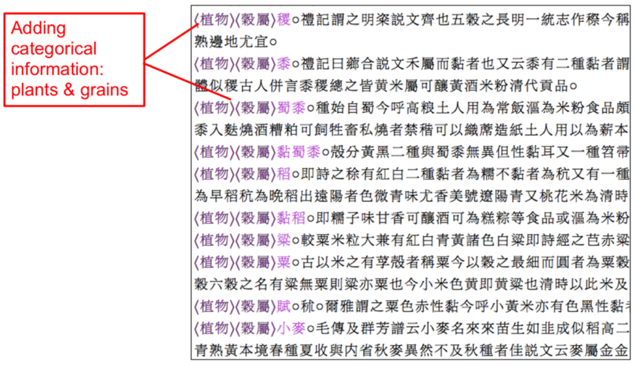
【1035】○○此為大宗○黏蜀黍○殼分黃黑二種與蜀黍無異但性黏耳又一種笤帚黍性黏穗長可製為帚○稻○即詩之稌有紅白二種黏者為糯不黏者為秔又有一種日秈比秔小早熟今人號○○○秈為早稻秔為晚稻出遠陽者色微青味尤香美號遂陽青又桃花米為清時貢品○黏稻○即糯子味甘香可釀酒可為糕粽等食品或漚為米粉最佳○粱○較粟米粒大兼有紅白青黃諸色白粱即詩經之芑赤粱即爾雅之虋○粟○古以米之有莩殼者稱粟今以穀之最細而圓者為粟穀曰飯穀米曰小米按周禮○○○九穀六穀之名有粱無粟則粱

Figure 2 A sample digital texts for "local products" for Figure 1. The categories are circled in blue color, while the names of products are circled in pink.



〈植物〉○
〈穀屬〉
○
稷○禮記謂之明粢說文齊也五穀之長明一統志作穄今稱糜子
黍○禮記日薌合說文禾屬而黏者也又云黍有二種黏者謂秫不
言黍稷總之皆黃米屬可釀黃酒米粉清代貢品○
蜀黍○種始自蜀今呼高粮土人用為常飯漚為米粉食品頗佳又
粕可飼牲畜私燒者禁稭可以織蓆造紙土人用以為薪本境土產
【1035】○○此為大宗○
黏蜀黍○殼分黃黑二種與蜀黍無異但性黏耳又一種笤帚黍性
稻○即詩之稌有紅白二種黏者為糯不黏者為秔又有一種日秈
稻出遠陽者色微青味尤香美號遼陽青又桃花米為清時貢品○
黏稻○即糯子味甘香可釀酒可為糕粽等食品或漚為米粉最佳○
粱○較粟米粒大兼有紅白青黃諸色白粱即詩經之芑赤粱即爾雅之虋○
粟○古以米之有莩殼者稱粟今以穀之最細而圓者為粟穀曰飯穀米曰小米按周禮○○○九穀六穀之名有
粱無粟則粱亦粟也今小米色黃即黃粱也清時以此米及穀草充○○○
賦○秫○爾雅謂之粟色赤性黏今呼小黃米亦有色黑性黏者○
小麥○毛傳及群芳譜云小麥名來來苗生如韭成似稻高二三尺實居殼中芒生殼上○○○生青熟黃本境春
種夏收與內省秋麥異然不及秋種者佳說文云麥屬金金旺而生○

Figure 3 Step 1 of turning texts into data: Break the text into records (rows) via the help of the Smart Regex tool.

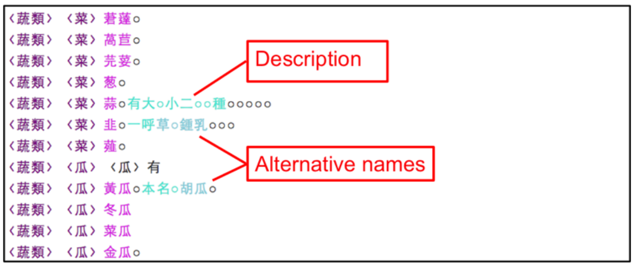


〈植物〉〈穀屬〉稷○禮記謂之明粢說文齊也五穀之長明一統志作穄今稱
熟邊地尤宜○
〈植物〉〈穀屬〉黍○禮記日薌合說文禾屬而黏者也又云黍有二種黏者謂
體似稷古人併言黍稷總之皆黃米屬可釀黃酒米粉清代貢品○
〈植物〉〈穀屬〉蜀黍○種始自蜀今呼高粮土人用為常飯漚為米粉食品頗
黍入麩燒酒糟粕可飼牲畜私燒者禁稭可以織蓆造紙土人用以為薪本
〈植物〉〈穀屬〉黏蜀黍○殼分黃黑二種與蜀黍無異但性黏耳又一種笤帚
〈植物〉〈穀屬〉稻○即詩之稌有紅白二種黏者為糯不黏者為秔又有一種
為早稻秔為晚稻出遠陽者色微青味尤香美號遼陽青又桃花米為清時
〈植物〉〈穀屬〉黏稻○即糯子味甘香可釀酒可為糕粽等食品或漚為米粉
〈植物〉〈穀屬〉粱○較粟米粒大兼有紅白青黃諸色白粱即詩經之芑赤粱
〈植物〉〈穀屬〉粟○古以米之有莩殼者稱粟今以穀之最細而圓者為粟穀
穀六穀之名有粱無粟則粱亦粟也今小米色黃即黃粱也清時以此米及
〈植物〉〈穀屬〉賦○秫○爾雅謂之粟色赤性黏今呼小黃米亦有色黑性黏者
〈植物〉〈穀屬〉小麥○毛傳及群芳譜云小麥名來來苗生如韭成似稻高二
青熟黃本境春種夏收與內省秋麥異然不及秋種者佳說文云麥屬金金

Figure 4 Step 2: Adding shared categorical information to each row.

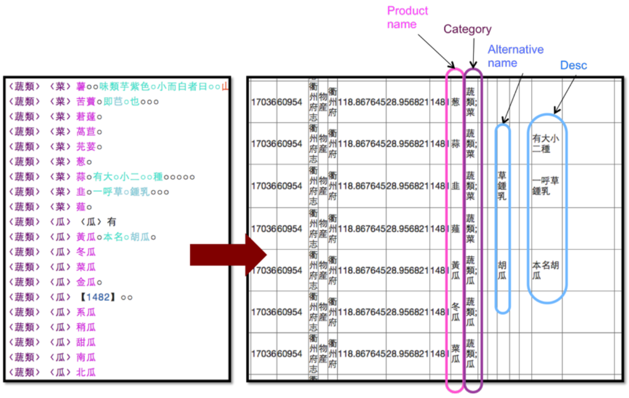
II. Compiling a global database from local records: A research data repository

Our extraction interface provides a semi-automatic way of transforming texts into data tables. Nevertheless, compiling a global database through this interface will still take a lot of time. We envision it to be a collaborative work among scholars with different research interests and specialties, and we eagerly need to be a way to give credits to scholars who contribute the data in the academic world. We found that the philosophy behind the open source software Dataverse, developed by Institute for Quantitative Social Sciences at Harvard University with the philosophy of "dedicated to sharing, archiving, and citing research data" (IQSS, 2015), matches our needs and will allow our project to meet the open access policy of the Max Planck Society and the "Berlin declaration".[2]



〈蔬類〉〈菜〉莙薘○
〈蔬類〉〈菜〉萵苣○
〈蔬類〉〈菜〉芫荽○
〈蔬類〉〈菜〉葱○
〈蔬類〉〈菜〉蒜○有大○小二○○種○○○○○
〈蔬類〉〈菜〉韭○一呼草○鍾乳○○○
〈蔬類〉〈菜〉薤○
〈蔬類〉〈瓜〉〈瓜〉有
〈蔬類〉〈瓜〉黃瓜○本名○胡瓜○
〈蔬類〉〈瓜〉冬瓜
〈蔬類〉〈瓜〉菜瓜
〈蔬類〉〈瓜〉金瓜○

Figure 5 Step 3: Tag further information that you want to capture.

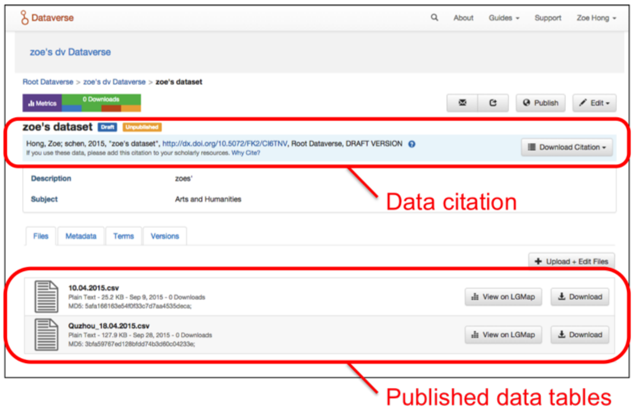


〈蔬類〉〈菜〉薯○○味類芋紫色○小而白者日○○山
〈蔬類〉〈菜〉苦蕒○即苣○也○○○
〈蔬類〉〈菜〉莙薘○
〈蔬類〉〈菜〉萵苣○
〈蔬類〉〈菜〉芫荽○
〈蔬類〉〈菜〉葱○
〈蔬類〉〈菜〉蒜○有大○小二○○種○○○○○
〈蔬類〉〈菜〉韭○一呼草○鍾乳○○○
〈蔬類〉〈菜〉薤○
〈蔬類〉〈瓜〉〈瓜〉有
〈蔬類〉〈瓜〉黃瓜○本名○胡瓜○
〈蔬類〉〈瓜〉冬瓜
〈蔬類〉〈瓜〉菜瓜
〈蔬類〉〈瓜〉金瓜○
〈蔬類〉〈瓜〉【1482】○○
〈蔬類〉〈瓜〉系瓜
〈蔬類〉〈瓜〉稍瓜
〈蔬類〉〈瓜〉甜瓜
〈蔬類〉〈瓜〉南瓜
〈蔬類〉〈瓜〉北瓜

17036	60954	衢州府志	物產	衢州府	118.86764	28.956821	148	葱	蔬類;菜			
17036	60954	衢州府志	物產	衢州府	118.86764	28.956821	148	蒜	蔬類;菜			有大小二種
17036	60954	衢州府志	物產	衢州府	118.86764	28.956821	148	韭	蔬類;菜	草鍾乳		一呼草鍾乳
17036	60954	衢州府志	物產	衢州府	118.86764	28.956821	148	薤	蔬類;菜			
17036	60954	衢州府志	物產	衢州府	118.86764	28.956821	148	黃瓜	蔬類;瓜	胡瓜		本名胡瓜
17036	60954	衢州府志	物產	衢州府	118.86764	28.956821	148	冬瓜	蔬類;瓜			
17036	60954	衢州府志	物產	衢州府	118.86764	28.956821	148	菜瓜	蔬類;瓜			

Figure 6 Step 4: Export the tagging result as a table.

After a text is tagged in our extraction interface, a user can publish the data to LGDataverse, a Dataverse instance we set up for this project. A citation link to the data along with the contributors will be shown on the LGDataverse page, urging any scholar who wishes to use the data to cite it in their publications (Figure 7).



zoe's dv Dataverse

Root Dataverse > zoe's dv Dataverse > zoe's dataset

zoe's dataset Draft Unpublished

Hong, Zoe; schen, 2015, "zoe's dataset", http://dx.doi.org/10.5072/FK2/CI6TNV, Root Dataverse, DRAFT VERSION
If you use these data, please add this citation to your scholarly resources. Why Cite?

Description zoes'
Subject Arts and Humanities

10.04.2015.csv
Plain Text - 25.2 KB - Sep 9, 2015 - 0 Downloads
MD5: 5afa166163e54f0f33c7d7aa4535deca;

Quzhou_18.04.2015.csv
Plain Text - 127.9 KB - Sep 28, 2015 - 0 Downloads
MD5: 3bfa59767ed128bfdd74b3d60c04233e;

Figure 7 A screenshot of LGDataverse.

We are still working on a way to aggregate the produced tables into one database in order to enable joint queries on records drawn from different gazetteers.

III. Computer assisted analyses on large scales

By transforming texts into data tables with predefined shared schema, it enables computers to easily process and analyze the data on large scales with better accuracy. We envision there can be multiple tools connected to our data, and scholars can choose which tools to use based on their research needs.

At the moment we are using the open source software PLATIN GeoBrowser (http://platin.mpiwg-berlin.mpg.de/) to create geospatial visualization for our data with one click away. PLATIN doesn't just display the geospatial distribution of the data on a map but also provides an animated timeline and user-defined pie charts, which we found very helpful for historians for preliminary analysis (Figure 8 and Figure 9).

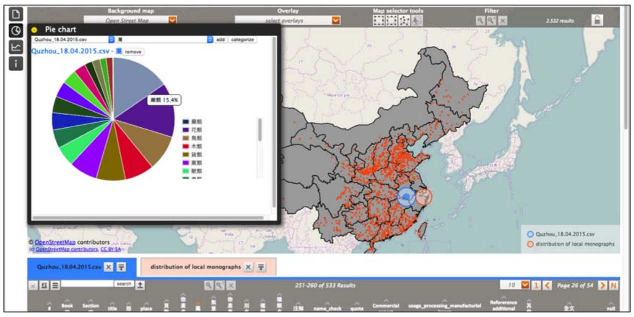


Figure 8 LGMap service (based on PLATIN) with Pie Chart function for analyzing categorical data.

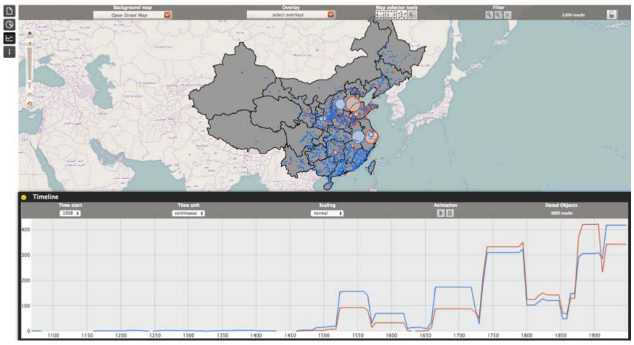


Figure 9 LGMap service with Timeline function showing temporal distribution of two datasets.

Concluding Remarks

This is an ongoing project. As a history of science study, we are collecting data from the chapter of local products in order to understand how local identities were constructed through the compilation of local products. In 2016, MPIWG will invite visiting scholars to use our digital tools to collect and analyze data of other topics from local gazetteers. This will also be a chance to further examine the model we proposed: from data collection from digital texts, aggregation of data in a research data repository, to digital tools for analysis and visualization.

Bibliography

Chang, S. (2015). Data Extraction and Analysis: The Taxonomy of Fauna and Flora of Taiwan Local Gazetteers in the Qing Dynasty. Unpublished paper presentation at: *Chinese Local Gazetteers workshop: Historical Methods and Computerized Data Collection and Analysis*, Berlin, Germany.

Chen, S.-P., Hsiang, J., Tu, H.-C. and Wu, M.-C. (2007). On Building a Full-Text Digital Library of Historical Documents. In Goh, D., Cao, T., Sølvberg, I. and Rasmussen, E. (eds.) *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*. Berlin: Springer, pp. 49-60.

Flanders, J. (n.d.). Women Writers Project. Retrieved November 1, 2015, from http://www.wwp.northeastern.edu/

Google Ngram Viewer. (n.d.). Retrieved November 1, 2015, from https://books.google.com/ngrams/

Institute for Quantitative Social Science (IQSS), Harvard University. *Dataverse Project* [Online]. Available at: http://dataverse.org/ (Accessed: November 1, 2015).

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Urbana, IL: University of Illinois Press.

Mitchell, A. (2015). Encoding China's Past: Computational Methods of Historical Analysis. In Unpublished paper presentation *Chinese Local Gazetteers workshop: Historical Methods and Computerized Data Collection and Analysis*, Berlin, Germany.

Pang, W.H., Cheng, H. and Chen, S.-P. (2014). From Text to Data: Extracting Posting Data from Chinese Local Gazetteers. In *the* 5th *International Conference of Digital Archives and Digital Humanities: Crossover and Transformation*, Taipei: National Taiwan University, pp. 93-116.

TEI: Text Encoding Initiative [Online]. Available at: http://www.tei-c.org/index.xml (accessed: November 1, 2015).

Wikipedia. *Regular expression* [Online]. Available at: http://en.wikipedia.org/wiki/Regular_expression (accessed: November 1, 2015).

Notes

[1] To name just a few of such digital local gazetteers project, there is first the *Chinese Local Gazetteers Database* (Zhongguo Fangzhii Ku) by a commercial vendor Erudition, second open access projects such as CTEXT.org with OCR-ed full texts of the gazetteers, and third national projects in China such as the Electronic Local Gazetteers by the Beijing National Library.

[2] Please see the declaration at http://openaccess.mpg.de/Berliner-Erklaerung.

Who's Doing What?: Examining The Relationships Among Subjectivity, Agency, and Syntax In The 19th Century Novel

Jonathan Yu Cheng
jonathan.cheng@huskers.unl.edu
University of Nebraska-Lincoln, United States of America

Gabrielle Kirilloff
gkirilloff@gmail.com
University of Nebraska-Lincoln, United States of America

Introduction

The relationship between action and identity is a significant element of understanding the way that characterization functions within literary works; many memorable characters are in part defined by their actions. This link between character and action raises the question of whether **specific** types of characters, or subjects, are consistently associated with certain types of action. Our project seeks to address this question by looking at the relationship between elements of a subject's identity and the actions associated with that subject. Our research builds off of work begun by the University of Nebraska-Literary Lab that explores the relationship between behavior and gender in the 19th century novel. The research begun by the Lab attempts to situate questions of gender and agency within the context of 19th century notions of propriety; is the Victorian valorization of passive women and active men reflected in novels from the period?

This project adds on to our initial foray into questions of gender; what is at stake is still very much a question of the allocation of agency. This avenue of research revolves around the question of when and why inanimate objects fill the subject position in sentences. This research also queries whether certain types of characters behave differently from others: what do kings do that peasants do not? Our project examines the agency associated with male, female, human, and non-human actors by studying the different types of verbs used in conjunction with different types of subjects. This research explores the question of whether or not certain types of subjects **behave** differently in our corpus, and if so, in what ways and to what effect.

Methodologies

The initial foray into the study of gender and genre performed by the University of Nebraska-Literary Lab relied on POS tagging and used an R programming script to extract the first pronoun that it encountered, along with the first verb that followed this pronoun, and entered each as a relationship into a data frame. The male pronouns "him," "his," "he," and "himself," and the female pronouns "she," "her," "hers," and "herself" were extracted. Thus, in the following sentence, the pronouns "she" would be extracted and grouped with the verb "walked."

> After dinner, she walked outside.

This approach was also our initial model for extracting non-human actors. For example, in the following sentence, we could similarly extract the pronoun "it" and the verb "howled."

> The wind was fierce; it howled into the night.

However, such an approach has several shortcomings. The first of which is that multiple verbs associated with a single subject in a sentence are not extracted. The second, is that this method only captures pronouns. Instances of personification, which often rely on nouns rather than pronouns, are ignored by this model. In order to solve these issues, we turned to the Stanford Dependency Parser, a tool that provides a representation of grammatical relations between words in a sentence. For example, in the sentence "The wind is dancing and howling," the parser would extract two subject verb pairs, "wind, dancing" and "wind, howling." The output looks as follows:

```
det(wind-2,The-1)
nsubj(dancing-4,wind-2)
nsubj(howling-6,wind-2)
aux(dancing-4,is-3)
root(ROOT-0,dancing-4)
cc(dancing-4,and-5)
conj:and(dancing-4, howling-6)
```

Using the parser allowed us to collect subjects that were not pronouns and allowed us to correctly associate multiple verbs with a single subject. It also allowed us to easily collect gender data, since we could simply collect any nsubj pair that contained a gendered pronoun.

While the parser does identify subject and verb pairs, it does not differentiate between human and non-human subjects. To differentiate between these subject types, we created a script that allows us to extract non-human agents and the verbs associated with them by ignoring sentences in which the subject is a gendered pronoun, a proper name, or a title. In performing our research, we realized that human subjects were indicated by either a pronoun (such as he), a proper name (such as Mary), or a title (such as the priest). If a subject did not fall into one of these three categories, the subject was most likely a non-human entity.

Ignoring nsubj groupings in which one of the words is a gendered pronoun was straightforward. In order to block proper names, we ran the Stanford Named Entity Recognizer on the texts in order to create a list of proper names from the corpus. We then ignored nsubj groupings that contained one of these names. Finally, in order to ignore titles, we created a dictionary of titles derived from vocabulary lists for non-native english speakers. These lists contained titles such as "captain," professions, such as "baker," terms signalling family relationships, such as "mother," and general terms for human agents, such as "girl." We then recorded each subject-verb relationship that did not contain one of these three categories into a data frame. However, in a separate script, we also used this list of titles to extract nsubjs that contain any of these titles. Our process allows us to use our program to assess the frequency of recurring syntactical relationships, essentially counting the number of times each verb is associated with male, female, human, and non-human subjects.

Observations

The initial results observed by the Nebraska-Literary Lab in their study on gender indicate that certain verbs were strongly associated with male characters while different verbs were strongly associated with female characters. Continuing this research, Matthew Jockers and Gabrielle Kirilloff confirmed these results in their work, which used the Stanford Dependency Parser in the manner discussed above. Jockers and Kirilloff found that a verb can be used to predict the gender of the pronoun associated with it, with 89% percent accuracy. Given the high degree of accuracy obtained from this analysis, we can conclude that within our corpus of 19th century fiction, authors chose to portray male and female characters differently by associating them with divergent groups of verbs. This result is not surprising, especially given the way in which ideas about proper behavior differed for males and females within 19th century society. However, this result still has several far-reaching implications, one of which is that "actions," or verbs, are in fact an important part of creating and determining character.

One of the shortcomings of the analysis on gendered pronouns and verbs is that it does not take into account other aspects of character identity. A princess and a witch may perform the same actions, but the implications are radically different. Similarly, certain types of characters may be associated with verbs typically associated with the opposite gender; though both are male, clerics and soldiers are no doubt associated with different actions. The data we extracted is a first step toward broadening this work; our extraction of specific subjects (such as wife, soldier, cleric) allow us to more closely look at character identity. In querying our data, our results thus far support the findings on gendered pronouns and verbs. For example, the verb "wept" was found to be strongly associated with female pronouns. In examining specific types of actors, we found that "women," "mothers," and "woman" were the most frequent actors associated with "weep," "weeping," and "weeps" respectively.

Future work

Our initial foray into our corpus has produced a wealth of data; at this stage our next step is to organize and query this data, asking more specific questions about the relationship between subjects and actions. For example, we have hypothesized that instances of objects performing actions occur more often in certain genres, specifically the Gothic. Over the coming months we intend to begin studying whether the actions associated with male, female, human, and non-human subjects are associated with specific genres. This type of analysis is challenging, largely because of the difficulties associated with collecting accurate Genre data. Genres are not rigid categories and many works participate in multiple genres. In addition to exploring the effects of genre, we also intend to more thoroughly examine the types of non-human and human agency we are extracting. Man-made objects, objects found in nature, animals, and supernatural beings are just a few of the types of non-human agency we have observed. We would like to begin exploring and categorizing these differences in an attempt to better understand our data. Because the Stanford Dependency Parser allows us to look closely at syntactic relationships, we also intend to expand our research to encompass the objects of actions, essentially asking, who is doing what **to whom**. This question has important implications for studies of gender and character identity.

Bibliography

Baylog, O., Dimmit, L., Heller, T., Kirilloff, G., Smith, S., Thomas, G., Warren, C. and Wehrwein, J. (2014). More than Custom has Pronounced Necessary: Exploring the Correlation between Gendered Verbs and Character in the 19th Century Novel, *UNL Digital Commons*.

Notes

[1] R is a statistical programming language often used in text analysis research and authorship attribution studies

[2] For a helpful discussion of the gender stereotypes that existing in the 19th century, please see: **Welter, B.** (1966). The Cult of True Womanhood: 1820-1860. *American Quarterly* **18**(2). **Clark, A.** (1995). *The Struggle for the Breeches: gender and the making of the British working class*. Berkeley, CA: University of California Press. **Gilbert, M. and Gubar, S.** (1979). *The madwoman in the attic: the woman writer and the nineteenth-century literary imagination*. New Haven, CT: Yale University Press. These works were influential in our understanding of 19th century notions of gender, behavior, and propriety.

[3] This hypothesis arose from both our own close reading of certain texts within our corpus and previous scholarship on the appearance and use of personification in the Gothic novel. For insight into the scholarly understanding of personification and the Gothic novel, please see the chapter on the Gothic novel in: **Parrinder, P., Nash, A. and Wilson, N.** (2015). *New Directions in the History of the Novel*. New York: St. Martin's.

Etymology Meets Linked Data. A Case Study In Turkic

Christian Chiarcos
christian.chiarcos@web.de
University of Frankfurt, Germany

Frank Abromeit
abromeit@unitybox.de
University of Frankfurt, Germany

Christian Fäth
chris.faeth.de@t-online.de
University of Frankfurt, Germany

Max Ionov
max.ionov@gmail.com
University of Frankfurt, Germany

Linking Etymological Dictionaries

When studying low-resource languages, historical documents or dialectal variation, researchers often face the problem that lexical resources are sparse, dated, or simply unavailable. At the moment, the problem is addressed by different initiatives to either aggregate language resources[1] in a central repository or to collect metadata about them[2]. The availability of this huge and diverse amount of material, often in different formats, and with a highly specialized focus on selected language varieties, poses the challenge how to access and search this wealth of information. Our project aims to address both aspects:

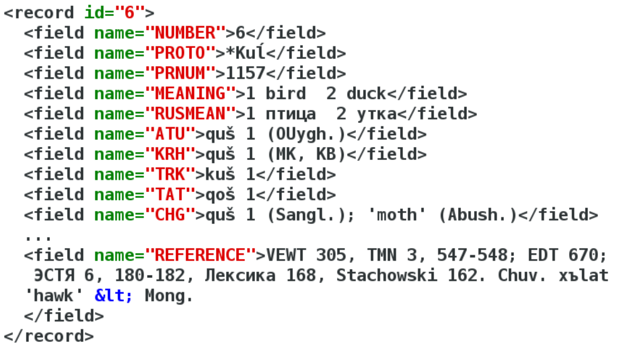
- **uniform access to lexical resources**. At the moment, most resources are distributed across different providers. Platforms to query or browse this data are available, but they use different representation formalisms and remain isolated from each other. We employ **Linked Data** to develop interoperable representations to access distributed resources in a uniform fashion.
- **search across multilingual resources**. We are not only interested in a specific language, but also, in related varieties: Much of the material we have is sparse, and we can address gaps in our lexical knowledge by consulting background information about form and meaning of possible cognates in other languages.

The project will implement search functionalities as web services and provide a prototypical web interface that allows to query Linked Data versions of open lexical resources. As a first step towards this goal, this paper addresses representation formalisms and data modelling, illustrated for an etymological dictionary of the Turkic language family.

Linked Open Data

Linked (Open) Data defines rules of best practice for publishing data on the web, and since (Chiarcos et al., 2012), these rules have been increasingly applied to language resources, giving rise to the **Linguistic Linked Open Data** (LLOD) cloud (Chiarcos et al., 2013)[3]. A *linguistically relevantresource* constitutes Linguistic Linked (Open) Data if (1) its elements are uniquely identifiably by means of **URIs**, (2) its URIs **resolve via HTTP**, (3) it can be accessed using **web standards** such as RDF and SPARQL, and (4) it includes **links** to other resources. It is Linguistic Linked Open Data (LLOD) if – in addition to these rules –, it is published under an **open** license. For etymological dictionaries, the capability to refer to and to search across distributed data sets (federation, dynamicity, ecosystem) in an interoperable way (representation, interoperability) allows to design novel, integrative approaches on accessing and using etymological databases, but only if common vocabularies and terms already established in the community are being used, re-used and extended. (Moran & Brümmer, 2013) established lemon (McCrae et al., 2011)[4] for representing etymological data. Inspired by the pre-lemon inventory (de Melo, 2014), we introduce lemon extensions for etymological relations, illustrated for the linked data edition of the Starling Turkic etymological dictionary. With further dictionaries for Turkic languages becoming available as a result of our project, these are linked with each other and with language resources from contact languages such as Mongolian, Iranian, Caucasian, Arabic, and Russian.

Turkic Etymology in Starling

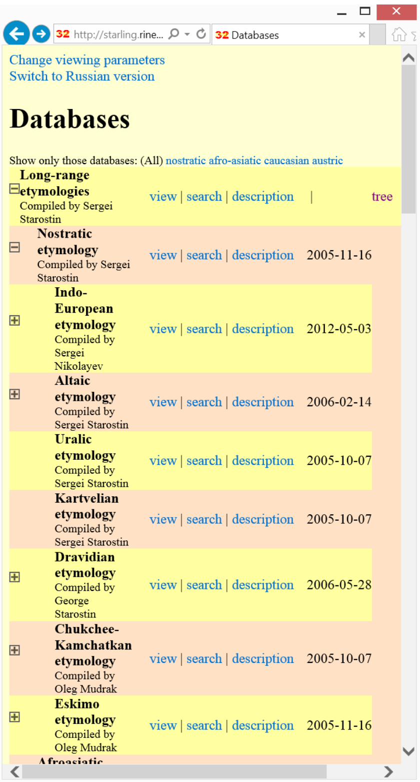


```
<record id="6">
  <field name="NUMBER">6</field>
  <field name="PROTO">*Kuĺ</field>
  <field name="PRNUM">1157</field>
  <field name="MEANING">1 bird  2 duck</field>
  <field name="RUSMEAN">1 птица  2 утка</field>
  <field name="ATU">quš 1 (OUygh.)</field>
  <field name="KRH">quš 1 (MK, KB)</field>
  <field name="TRK">kuš 1</field>
  <field name="TAT">qoš 1</field>
  <field name="CHG">quš 1 (Sangl.); 'moth' (Abush.)</field>
  ...
  <field name="REFERENCE">VEWT 305, TMN 3, 547-548; EDT 670;
   ЭСТЯ 6, 180-182, Лексика 168, Stachowski 162. Chuv. хъlat
  'hawk' &lt; Mong.
  </field>
</record>
```

Fig. 1: XML snippet

The **Tower of Babel (Starling)**[5] is a web portal on historical and comparative linguistics (Starostin, 2010), widely used in academia to publish etymological datasets over the internet. Starling allows exploring its dictionaries by means of faceted browsing using a coarse-grained phylogenetic tree (Fig. 2.a). We illustrate its data structures for the Turkic Etymological Dictionary (Dybo et al., 2012) with an example result for the query **meaning="bird"** (Fig.2.b).

Following the **Proto-Turkic** root, we find a cross-reference to the Altaic dictionary, and the **meaning** (sense) of the proto-form in English and Russian. The following entries pertain to **cognates** in different Turkic languages: They provide complex information including one or multiple **forms**, **co-indexed** with the meaning field, and optionally augmented with additional gloss (e.g., 'moth' for Middle Turkic/Chagatai), bibliography (as a hyperlink, Fig. 3) or additional comments (e.g., < Az. for Halaj). We used an XML export of the Starling data (Fig. 1) to create RDF and (by converting cross-references) Linked Data.



Change viewing parameters
Switch to Russian version

Databases

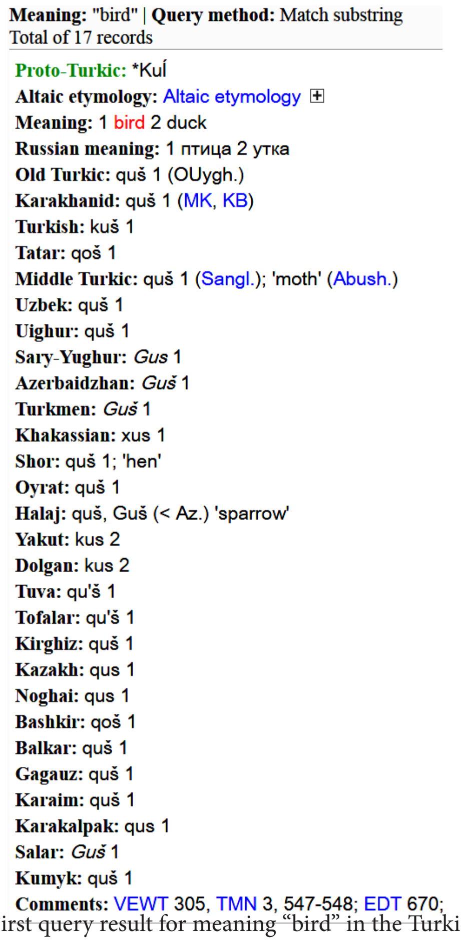
Show only those databases: (All) nostratic afro-asiatic caucasian austric

⊟	**Long-range etymologies** Compiled by Sergei Starostin	view | search | description |	tree
⊟	**Nostratic etymology** Compiled by Sergei Starostin	view | search | description	2005-11-16
⊞	**Indo-European etymology** Compiled by Sergei Nikolayev	view | search | description	2012-05-03
⊞	**Altaic etymology** Compiled by Sergei Starostin	view | search | description	2006-02-14
	Uralic etymology Compiled by Sergei Starostin	view | search | description	2005-10-07
	Kartvelian etymology Compiled by Sergei Starostin	view | search | description	2005-10-07
⊞	**Dravidian etymology** Compiled by George Starostin	view | search | description	2006-05-28
⊞	**Chukchee-Kamchatkan etymology** Compiled by Oleg Mudrak	view | search | description	2005-10-07
⊞	**Eskimo etymology** Compiled by Oleg Mudrak	view | search | description	2005-11-16
	Afroasiatic		

Fig. 2a: Starling phylogenetic tree for faceted browsing

Data Model for the Turkic Etymology

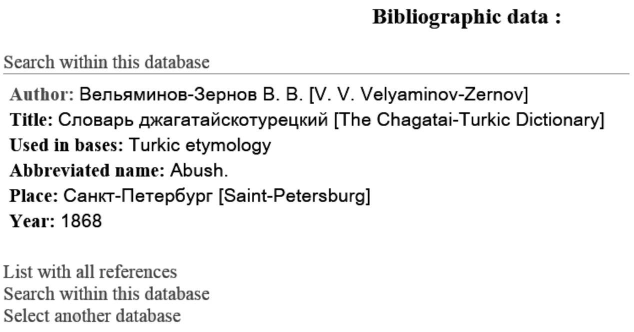
Following LLOD conventions, we employ the Ontolex/Lemon vocabulary (McCrae et al., 2011)[6] as shown in Fig. 4. Originally developed to add linguistic information to existing ontologies, Lemon evolved into a de-facto standard to represent lexical resources as LLOD. Here, we focus on Lemon extensions to represent etymological cognates: Etymological relations involve a relationship on the level of meaning (sense) and on the level of form, and thus require a novel property between one **LexicalEntry** and another. Between etymological cognates, it is not always clear whether one was the source of the other, or a more indirect relation holds. To express a generic etymological link without additional directionality information, we introduce the property **lemonet:cognate**. If source and target are known, a subproperty **lemonet:derivedFrom** is introduced. Similar to **lemonet:cognate**, it is transitive, but it is not symmetric. Distinguishing **lemonet:cognate** and **lemonet:derivedFrom** follows de Melo's apparent directionality differentiation. Here, however, we provide a formal definition as a (minimal) extension of Lemon following (Chiarcos & Sukhareva, 2014) which supports inferring general cognate relations by subsumption and transitive/symmetric closure. In the Starling data, the directionality of etymological links is generally known, so we represent etymological relations with **lemonet:derivedFrom** between lexical entries from different Lemon lexicons. By subsumption inference, transitivity and symmetry of its superproperty, **lemonet:cognate** relations can be inferred automatically between all language-specific forms.



Meaning: "bird" | **Query method:** Match substring
Total of 17 records

Proto-Turkic: *Kuĺ
Altaic etymology: Altaic etymology ⊞
Meaning: 1 bird 2 duck
Russian meaning: 1 птица 2 утка
Old Turkic: quš 1 (OUygh.)
Karakhanid: quš 1 (MK, KB)
Turkish: kuš 1
Tatar: qoš 1
Middle Turkic: quš 1 (Sangl.); 'moth' (Abush.)
Uzbek: quš 1
Uighur: quš 1
Sary-Yughur: *Gus* 1
Azerbaidzhan: *Guš* 1
Turkmen: *Guš* 1
Khakassian: xus 1
Shor: quš 1; 'hen'
Oyrat: quš 1
Halaj: quš, Guš (< Az.) 'sparrow'
Yakut: kus 2
Dolgan: kus 2
Tuva: qu'š 1
Tofalar: qu'š 1
Kirghiz: quš 1
Kazakh: qus 1
Noghai: qus 1
Bashkir: qoš 1
Balkar: quš 1
Gagauz: quš 1
Karaim: quš 1
Karakalpak: qus 1
Salar: *Guš* 1
Kumyk: quš 1
Comments: VEWT 305, TMN 3, 547-548; EDT 670;

Fig. 2b: First query result for meaning "bird" in the Turkic etymological dictionary



Bibliographic data :

Search within this database

Author: Вельяминов-Зернов В. В. [V. V. Velyaminov-Zernov]
Title: Словарь джагатайскотурецкий [The Chagatai-Turkic Dictionary]
Used in bases: Turkic etymology
Abbreviated name: Abush.
Place: Санкт-Петербург [Saint-Petersburg]
Year: 1868

List with all references
Search within this database
Select another database

Fig. 3: Bibliographic information for **Abush**

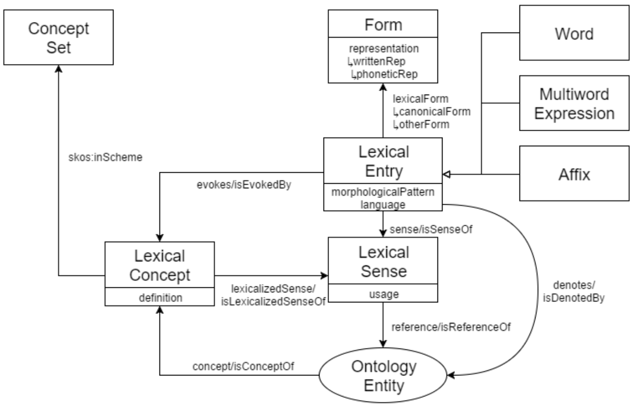


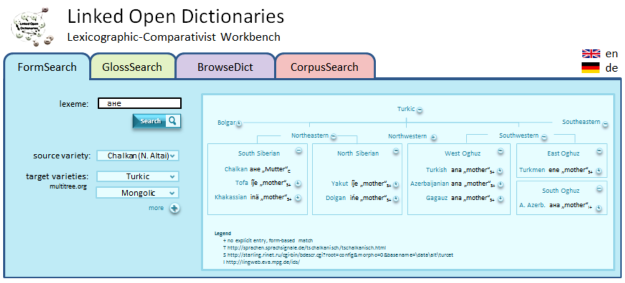
Fig. 4: Lemon-Core (Ontolex) module

Applications

The **Comparative-Lexicographical Workbench** (Fig. 5) will provide novel search functionalities extending the functionality of existing platforms, form-based search and a gloss-(meaning-) based search, currently applied to the Turkic language family and its contact languages.

- **gloss-(meaning-)based search**. Dictionary lemmas are complemented with a gloss paraphrasing their meaning. Linked Data allows transitive search over sequences of bilingual dictionaries (e.g., Kazakh-Russian-English).
- **form-based search**. Given a lexeme in a particular language, say, Kazakh, and a set of related languages, say, the Turkish languages in general, the system will retrieve phonologically similar lexemes for the respective target languages.

Both search functionalities aim to detect candidate cognates. The data provided by Starling represents a gold standard, but can also be directly integrated into the search process: In Fig. 5, we query for Chalkan **ана** and possible cognates from Turkic (as an inherited word) or Mongolic (as a possible source of loan words). The results are organized according to the taxonomic status of the varieties in *www.multitree.org*. They include a gloss from a Chalkan dictionary (marked by subscript C), but in addition provide form-based matches (subscript +) from the Starling dictionaries (S), e.g., with Turkish **ana** and its etymologically corresponding forms, etc.



lexeme: аhе

source variety: Chalkan (N. Altai)

target varieties: Turkic

multitree.org

Mongolic

Chalkan ане „Mutter"$_C$

Tofa ĳe „mother"$_{S+}$

Khakassian inä „mother"$_{S+}$

Yakut ĳe „mother"$_{S+}$

Dolgan íńe „mother"$_{S+}$

Turkish ana „mother"$_{S+}$

Azerbaijanian ana „mother"$_{S+}$

Gagauz ana „mother"$_{S+}$

Turkmen ene „mother"$_{S+}$

A. Azerb. ана „mother"$_{I+}$

Legend
+ no explicit entry, form-based match
T http://sprachen.sprachsignale.de/tschalkanisch/tschalkanisch.html
S http://starling.rinet.ru/cgi-bin/bdescr.cgi?root=config&morpho=0&basename=\data\alt\turcet
I http://lingweb.eva.mpg.de/ids/

Fig. 5: Design study: Form-based search in the Comparative-Lexicographical Workbench

Summary

We described preliminary steps towards the development of a Comparative-Lexicographical Workbench that uses Linked Data formalisms to retrieve cognates as given in etymological dictionaries as well as to automatically identify cognate candidates from different languages (which are similar in form and meaning). In our presentation, both will be illustrated for the Turkic language family, and we will show how both aspects complement each other.

Bibliography

Chiarcos, C., Nordhoff, S. and Hellmann, S. (2012). *Linked Data in Linguistics*. Berlin: Springer.

Chiarcos, C., Cimiano, P., Declerck, T. and McCrae, J. (2013). Linguistic linked open data (llod). *Proc. 2nd Workshop on Linked Data in Linguistics (LDL-2013)*, Pisa, Italy, pp. 1-11.

Chiarcos, C. and Sukhareva, M. (2014). Linking Etymological Databases. *Proc. 3rd Workshop on Linked Data in Linguistics* (LDL-2014), Reykjavik, pp. 41–49.

De Melo, G. (2014). Etymological Wordnet: Tracing the history of words. *Proc. LREC 2014*, Reykjavik.

Dybo, A. V., Starostin, S. A. and Mudrak, O. A. (2012). *Etymological Dictionary of the Altaic Languages*. Brill, Leiden.

McCrae, J., Spohr, D. and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *Proc. 8th Extended Semantic Web Conference (ESWC-2011)*, Heraklion, Crete, pp. 245–59.

Moran, S. and Brümmer, M. (2013). Lemon-aid: Using Lemon to aid quantitative historical linguistic analysis. *Proc. 2nd Workshop on Linked Data in Linguistics (LDL-2013)*, Pisa, Italy.

Starostin, G. (2010). Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship*, **3**: 79–117.

Notes

[1] e.g. ELRA http://www.elra.info

[2] e.g. OLAC http://www.language-archives.org/

[3] e.g. http://linguistics.okfn.org/, http://linguistic-lod.org

[4] http://lemon-model.net

[5] http://starling.rinet.ru/

[6] http://www.w3.org/community/ontolex/wiki/Final_Model_Specification

#ww1. The Great War on Twitter

Frédéric Clavert
frederic@clavert.net
Université de Lausanne, Switzerland

Since 2014, some of the countries that were formerly belligerent of the Great War – most particularly France and UK – have organised a series of commemorations of the First World War, known as the 'Centenaire' (France) or the 'Centenary' (UK). We can assume that there is a strong link – that cannot let a historian indifferent – between those commemorations, collective memory and historical studies.

Though studies about collective memory are numerous since the famous works of the French sociologist Maurice Halbwachs (Halbwachs, 1950), few of them are examining how collective memories are being expressed – maybe even transformed – on social networks on-line.

In the case of the Centenary of the First World War, a set of questions can be asked: What is the on-line echo of the commemoration of the centenary of the 1st World War? What is the behaviour of Memorial/Heritage Institutions about the 1st World War on Twitter? How do they transmit information about the Centenary? Is there an influence of the English predominance on Twitter about the Centenary on how non-english-speaking twitter accounts are considering the 1st World War? Are there specific subjects that are discussed on-line? Which 'temporalities' are present in tweets when Twitter users speak about the Great War on-line?

Though we are not yet able to respond to all those questions, we'll use our database of tweets in order to answer them at least partially.

Indeed, since the 1st April 2014, around 1.5 millions of tweets containing a hashtag (keyword) linked to the 1st World War were written by over 350 000 Twitter accounts in several languages (mainly English and French). Twitter is a good field to analyse relationships between history and collective memory, memorial institutions and citizens, historians and a wide non-academic audience. We started to explore this database (which is still expanding): we intend to show how a historian can collect, analyse and interpret those tweets, using Digital Humanities methodologies and software in order to answer questions about collective memory of the First World War online.

Tools and Methodologies

We are using 140dev, a PHP open source script within a LAMP environment to collect tweets through the Twitter streaming API. The tweets are then stored in a MySQL database. Diverse information (tweets and their metadata, hashtags, user information, mentions, retweets) about those tweets can easily be extracted through SQL queries. Those queries can also be used to extract different kind of relations: between tweets, between Twitter users or even between hashtags (*ie* if a Twitter user mentioned or retweeted another twitter user, if two users wrote the same hashtags, etc). Concerning privacy, we respect the Twitter API Terms.

To analyse tweets, we are using mainly two sets of methodologies/software: social network analysis and network visualisations (with Gephi: mention, retweets or hashtags are considered a link); text analysis through the theory of the *mondes lexicaux* (Reinert, 1993) as it is implemented in the IRaMuTeQ software (Ratinaud and Dejean, 2009) . The combination of both tools and methodologies has been described by (Smyrnaios and Ratinaud, 2014). IRaMuTeQ, thanks to time-stamped metadata, can also help us working on temporalities. Indeed, clusters that are defined by this software can be projected in time: we can know, day-by-day, the most used kind of tweets. It helped us, for instance, finding that French fallen soldiers are not described with the same words the 11th of November in comparison to the rest of the year.

The methodologies and tools that remain to be found for this research concern temporalities – even if IRaMuTeQ has helped us answer some question on time. There are several temporalities that are expressed in this corpus: the constant feed of information that is the nature of Twitter; the temporality of each twitter user; the temporality of the Centenary (which is different from one country to the other, and from the Great War temporality); and the temporality of the War itself.

First results

Language

English is overwhelmingly present in this corpus. Around 10% only of the collected tweets are not in English. Among those 10%, French is largely in majority and German almost absent, even though German hashtags are collected. The fact that Twitter is an English-based social network does not explain fully this disequilibrium between English and other languages. The Memorial institutions' communication policies on Twitter are better factors to explain it.

The decentralized communication policy of British memorial institutions (the BBC and all its Twitter accounts or the Imperial War Museum for instance) is obviously more efficient than the French centralized communication policy of the *Mission du centenaire*. French WW1-related museums do not have Twitter accounts or do have one but do not follow twitter implicit rules such as the use of a general hashtag like #ww1 or the French #pgm.

British and French are not commemorating WW1 the same way

The most striking difference between the French corpus and the English one is the fact that both linguistic areas do not commemorate the Great War the same way. There are two major differences between both countries:

- French are mainly remembering the soldiers (*Poilus*). British citizens are remembering soldiers, but also battles.
- The French are focusing on the end of the war, the Armistice, on the 11th November. The British are focusing on the way they entered the war.

English public history and French history amateurs

Thanks to the Network visualisations, this corpus also helps understand how public history is present in Britain, in contrary to France where it just begins to appear. The presence of amateurs of history in the French corpus also shows that French historians are not on twitter, in contrary to amateurs who, next to the *Mission du Centenaire,* are structuring discussions about the First World War on Twitter.

Conclusion

Comparing multilingual corpora

To compare our two main corpora (the French one and the English one) that can be extracted from the database, we had to use the two main pieces of software the same way on both corpora and then to 'humanly' compare the results. We could not find any tools able to compare two corpora that are in different languages.

Distant reading / Close reading

This research project shows that, for historians, it is still important to keep a direct link with each single primary source, as some information can be learned from the interpretation of single tweets. Though methods used in this research are dealing with Franco Moretti's notion of *distant reading* (Moretti, 2007), it proved strategic to be able to go back to every single tweet. The software used, if metadata are kept all along their use, allow this.

Twitter and the rest of the web

Why Twitter? The fact that the Twitter API, though sometimes very unstable, is very convenient to use is one of the criteria of this choice. Is it really pertinent in terms of research? Shouldn't we have broader sources? How to extrapolate the project's results to other on-line social networks? Last but not least, the difficulty to anticipate hashtags to be collected might introduce biases in our research.

Future of this research project

The question of 'temporalities' and their imbrications (the temporality of Twitter / the temporality of users / the temporality of the commemorations / the temporality of the First World War itself) should be the next step of this research. But, as it will require the use of Named Entity Recognition, extending our research to places will be possible as well.

Bibliography

Halbwachs, M. (1950). *La mémoire collective*, Paris: Albin Michel.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.

Ratinaud, P. and Dejean, S. (2009). IRaMuTeQ: Implémentation de la methode ALCESTE d'analyse de texte dans un logiciel libre [Implementation of the ALCESTE method of text analysis in an open-source software]. *Presentation*. Available at: http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf (accessed 4 March 2016).

Reinert, M. (1993). Les 'mondes lexicaux' et leur 'logique' à travers l'analyse statistique d'un corpus de récits de cauchemars. *Language et Société*, **66**(1): 5–39. doi:10.3406/lsoc.1993.2632

Smyrnaios, N. and Ratinaud, P. (2014). Comment articuler analyse des réseaux et des discours sur Twitter. *tic and société*, 7(2). doi:10.4000/ticetsociete.1578

Notes

[1] http://140dev.com/ (accessed 4 March 2016).

[2] http://www.iramuteq.org/ (accessed 4 March 2016) - Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. IRaMuTeQ is a free software based on python and R. It is available in French, English, German and Spanish (interface and analyses).

[3] IRaMuTeQ works in dividing the corpus in small segments of text (around 40 words). In our case each segment is a tweet and each tweet is also a text.

A Method for Record Linkage with Sparse Historical Data

Giovanni Colavizza
giovanni.colavizza@epfl.ch
EPFL, Switzerland

Maud Ehrmann
maud.ehrmann@epfl.ch
EPFL, Switzerland

Yannick Rochat
yannick.rochat@epfl.ch
EPFL, Switzerland

Introduction

Massive digitization of archival material, coupled with automatic document processing techniques and data visualisation tools offers great opportunities for reconstructing and exploring the past. Unprecedented wealth of historical data (e.g. names of persons, places, transaction records) can indeed be gathered through the transcription and annotation of digitized documents and thereby foster large-scale studies of past societies. Yet, the transformation of hand-written documents into well-represented, structured and connected data is not straightforward and requires several processing steps. In this regard, a key issue is entity record linkage, a process aiming at linking different mentions in texts which refer to the same entity. Also known as entity disambiguation, record linkage is essential in that it allows to identify genuine individuals, to aggregate multi-source information about single entities, and to reconstruct networks across documents and document series.

In this paper we present an approach to automatically identify coreferential entity mentions of type *Person* in a data set derived from Venetian apprenticeship contracts from the early modern period (16th-18th c.). Taking advantage of a manually annotated sub-part of the document series, we compute distances between pairs of mentions, combining various similarity measures based on (sparse) context information and person attributes.

Task Definition

Major challenges when dealing with people-related data are homographic person names referring to different persons as well as the existence of name variants referring to the same person. These are well-known issues in the field of Natural Language Processing for which various approaches have been devised, first via mention clustering (Mann and Yarowsky 2003; Artiles et al. 2008), more recently via linking to a knowledge base (Ji and Grishman 2011; Shen et al 2015).

In the context of historical data, dealing with person name ambiguity is all the more difficult since data is inherently sparse and uncertain (resulting in poor mention context) and since knowledge bases such as DBpedia (Lehmann et al 2013) contain very little about past average laypersons (resulting in poor entity context). It is however an essential step prior to any historical data analysis (Bloothooft et al 2015), which we address as part of the *Garzoni* project. This project aims at studying apprenticeship in early modern Venice by extracting information from archival material. Part of this material have been manually annotated, including mention links towards unique entities. Starting from a subset of the current data, we present a method for person record linkage, with the objective to complement its disambiguation coverage and to bootstrap a system to better automate entity disambiguation during annotation, in an active learning fashion.

The *Accordi dei Garzoni*

The *Accordi dei Garzoni* is a document series from the State Archives of Venice which originates from the activity of the *Giustizia Vecchia* magistracy. This judicial authority was in charge of registering apprenticeship contracts in order to protect young people while they were trained and/or providing domestic services (Bellavitis 2006). As a result of this regulation, information for much of apprenticeship arrangements got centralized, today reflected in a dense archival series.

The *Accordi* consists of about 55,000 contracts registered from 1575 until 1772. Each contract features an apprentice, a master and often a guarantor, sometimes two. A sample of 11,000 contracts have been manually annotated and the resulting data is stored in an RDF triple store. For each person mentioned in a contract, annotators created a *person mention* and, importantly, linked it to a *person entity*. They did so either by selecting an already existing entity in the database or by creating a new one. Given the difficulty of this task, only a limited number of entities have been disambiguated; the annotated dataset can therefore be considered as correct but not exhaustive.

The present work considers annotated documents from the period 1586-1600, for which statistics about contracts and entity/mention ratio are shown in Table 1. We use a subset of this dataset (bolded line in the table) as a *golden* set for our experiments.

Count	**whole period**	**1586-1600**
# annotated contracts	11,525	2,687
# mentions	31,952	7,589

# entities	26,641	6,599
# entities with # mention > 1	1793	382
avg mention per entity	1.09	1.08
avg mention per entity with # mention > 1	2.44	2.38

Table 1. Entity-Mention stastistical profile for the whole vs. selected period

Approach

Given a set of mentions, our objective is to estimate the likelihood that two mentions refer to the same entity. We represent each mention by a vector of features and compare them pairwise using various similarity measures. The list of selected features at mention and contract levels are presented in Table 2 and 3 respectively.

Feature	Variable Type
first name	string
surname	string
patronymic	string
gender	categorical
age	integer
profession	categorical
geographical origins	string

Table 2. Mention-level features

Feature	Variable Type
workshop toponym	string
workshop parish	string
workshop sestriere1	string
workshop insi gna	string
contract year	integer
contract duration	string
master profession	categorical

Table 3. Contract-level features

With respect to our dataset and features, several points should be emphasized. First, data sparsity: it is common for a mention to have just a few features. Second, features are not evenly sparse (cf. Figure 1) and do not contribute equally to a possible linkage. Core features such as *name*, *surname*, *patronymic*, *gender* and *profession* must strongly correspond in order to consider a link as reliable. On the other hand, rare features such as *workshop insigna* can be very informative when shared by two mentions and should also be valued by the linkage algorithm. Finally, features are dependent, particularly on the role of the person (e.g. age indicated only for apprentices).

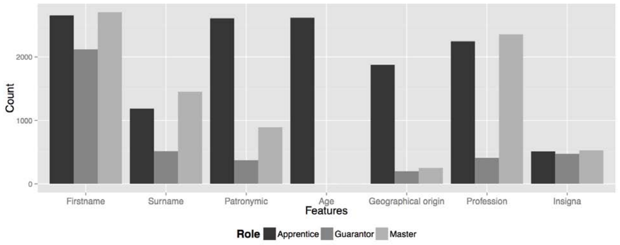


Figure 1. Distribution of features by role

We construct three matrices of size $N \times N$, where N is the number of mentions in the dataset. The first matrix Φ, the *feature matrix*, stores similarity scores of mentions pairwise. Scores are computed using measures over features as follows:

- *year of contract*: the feature-score is measured via distance and diminishing returns. Each year of distance between 1 and 15 and between 15 and 30 decreases an initial score of 1 by 0.01 and 0.025 respectively, with a definitive cut-off after 30. For example, two contracts from 1590 and 1594 have a score of 0.96.
- *age*: similarly as per year, each year of distance of the difference between two ages decreases an initial score.
- *gender* and *profession*: the feature-scores are based on exact matches.
- *name*, *surname*, *patronymic* and *workshop toponym*: the feature-score is based on the Deverau-Levenshtein string metric (Cohen et al, 2003). For example, *Polo* and *Pollo* have a similarity measure of 0.95.
- *geographical origins* and *insigna*: the feature-score is based on a token-based variant of the Jaro-Winkler metric. For example, *Friulano* and *del Friuli* have a similarity measure of 0.82.

The score of each pair is stored in Φ: it is the L2 norm of the resulting feature-score vector.

The second matrix Γ, the *combination matrix*, stores values that indicate whether a pair of mentions shares similar feature combinations or not. To build such matrix, we leverage the *golden* set and identify combinations of features which produced a linkage on a role-by-role basis (e.g. master-master or guarantor-master). Features are considered activated when their feature-score is equal or above 0.84[2] and we filter out combinations occurring once. The score of a mention pair in Γ is 1.0 if the combination of activated features is valid for the given role pair; 0.5 if the role pair does not match but the combination is valid; 0.0 otherwise. This matrix accounts for feature dependencies and the different ways to name a person with respect to his/her role.

The third matrix Δ, the *filtering matrix*, scores mention pairs according to the number of activated core features (1.0 if 3 + features – out of 5 – are activated, 0.0 otherwise[3]).

Given the three matrices, we normalize them and consider the following function to compute the similarity score of a mention pair *p*:

$$S(p) = \delta_p[\lambda\pi_p + (1-\lambda)\gamma_p]$$

where δ_p is a boolean parameter taken from Δ which activates the filter over core features for pair *p*; π_p is the feature score taken from Φ; γ_p is the combination score from Γ; and λ is a parameter giving priority over vector features or combinations of features. $\delta \in \{0, 1\}$ and $0 \leq \lambda, \pi, \gamma \leq 1$. This function allows us to adjust the different parameters: core vs sparse features (δ), feature scores (π) and feature combinations (γ).

Evaluation

We evaluate our approach in terms of coverage and precision. With respect to coverage, we verify our method over 100 thresholds from 0.99 to 0.0. For each threshold, we compare linkage curves as the percentage of links obtained over the total possible against the coverage of the *golden* set. Precision is based on manual annotation of 50 randomly selected linkages.

Both procedures are repeated with λ ∈ {0.1, 0.5, 0.9} and δ activated or not, for a total of 6 configurations. The objective is to understand the individual contributions of the three components to our function.

Results and Discussion

Results for the first and second evaluation procedure are presented in Figure 2 and Table 4 (resp.). Highest precision (0.61 and 0.3 in Table 4) is obtained with a balance between feature combinations and feature scores (λ = 0.5). δ proves very useful for filtering the input space (from 28,7M possible pairs to 44,2K), and lowers the number of false positives, especially for links between apprentices (cf. line 'w-o AA' in Table 4). The combination of the two (filtered input space and equal weights) provides the best results, especially for masters and guarantors. Linkage curves can be explained similarly: low λ entails a step-like curve (three steps at 0.0, 0.5 and 1.0), while high λ creates a Gaussian over the disambiguation space.

	active			δ not active		
	λ = 0.1	λ = 0.5	λ = 0.9	λ = 0.1	λ = 0.5	λ = 0.9
All	0.21	0.3	0.21	0.0	0.26	0.15
without A-A	0.22	0.61	0.22	0.0	0.48	0.67

Table 4. Precision with threshold >= 0.9 (* means not-significant statistics

This confirms that a balanced approach might be the best solution in a setting where data is sparse (high λ), the *golden* set is present but of limited coverage (low λ), and some prior assumptions on the required features can be made (δ). As shown in Figure 3, the graphs with λ = 0.5 and δ = 1.0 collapse more gradually, providing the widest effective linkage space to explore. Eventually, results also suggest to proceed in an active learning fashion, where the system learns iteratively with new data as part of the *golden* set.

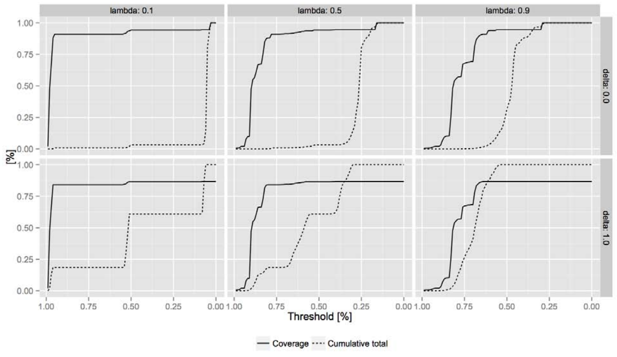


Figure 2. Linkage curves for the 6 parameters settings, over thresholds

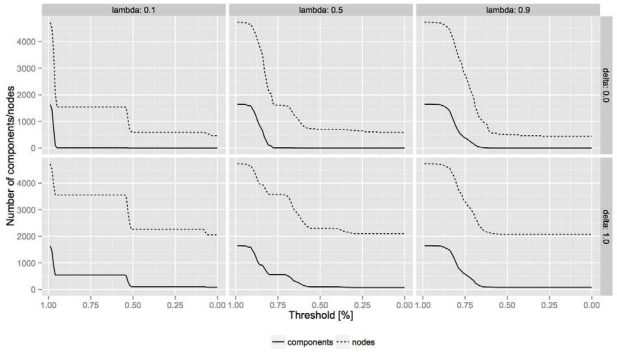


Figure 3. Graph properties for the 6 parameter settings, over thresholds

Finally, in order to further motivate our work, Figure 4 shows the largest components of the deduced social network with and without automatic disambiguation. The linkage method has the nice property of enlarging small components before gradually connecting them.

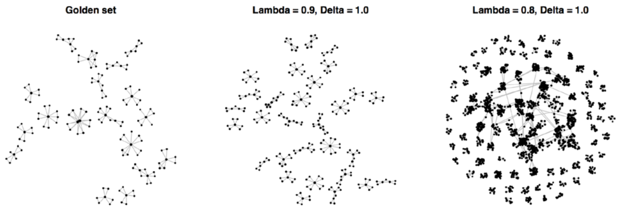


Figure 4. Largest components of social networks from golden set (left-most) and from disambiguated datasets (center and right-most)

Conclusion and Future Work

This paper presented a system to perform record linkage over mentions of persons from sparse historical data. It deals with different constraints such as data sparsity and limited prior knowledge. We plan to apply the system to different datasets and to integrate it into a transcription and annotation interface, in order to use it for live, aided record linkage.

Bibliography

Artiles, J., Sekine, S. and Gonzalo, J. (2008). Web people search: results of the first evaluation and the plan for the second. *Proceedings of the 17th international conference on World Wide Web* pp. 1071–72. ACM.

Bellavitis, A. (2006). Apprentissages masculins, apprentissages fminins venise au XVIe siècle. *Histoire Urbaine*, pp. 49–73.

Bloothooft, G., et al., (2015). *Population Reconstruction*. Springer.

Cohen, W., Ravikumar, P. and Fienberg, S. (2003). A comparison of string metrics for matching names and records. *KDD workshop on data cleaning and object consolidation*, **3**: 73–78.

Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume* **1**: 1148–58. Association for Computational Linguistics.

Kleanthi, G., et al. (2015). Record linkage in medieval and early modern text. *Population Reconstruction*, pp. 173–95. Springer.

Lehmann, J., et al. (2013). DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal.*

Mann, G. S. and Yarowsky, D. (2003). Unsupervised personal name disambiguation. *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, vol. **4**(3): 33–40.

Shen, W., Wang, J. and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering*, **27**(2): 443–60.

Notes

[1] There are 6 sestrieri in Venice, i.e. groups of contiguous parishes.

[2] It has been shown in comparable settings that edit distance with cut-off at distance 3 (which for us is distance > 0.85) provides good results (Kleanthi et al. 2015).

[3] Features are activated when their similarity is above 0.84.

Metacanon.org: Digital Humanities and the Canon

Nathaniel Allen Conroy
nathaniel_conroy@brown.edu
Brown University, United States of America

Can the digital humanities offer an alternative to traditional modes of canon formation? This paper argues that quantitative methods can both enrich our understanding of the way canons are formed and help us create more flexible and interactive alternatives to traditional canons. Over the last year, I've built a tool for generating dynamic literary canons and placed it in public beta at http://www.metacanon.org. Metacanon measures the canonicity of literary works by calculating the number of times they are mentioned in scholarly journals and using an algorithm to assign a uniform score to each work based on this data. While this certainly does not amount to a measure of aesthetic value or "greatness," it does offer a concise snapshot of the body of literary works that are most discussed by scholars. Currently, it only covers twentieth century American fiction, but future versions will be expanded to include other genres, periods, and nationalities. For scholars, this will provide a tool for quickly measuring the relative centrality or obscurity of particular works as well as a tool for measuring how canons change over time. For students and the general public, it will offer a far more inclusive, flexible, and interactive alternative to the fairly predictable greatest books lists that currently act as arbiters of literary value outside of academic circles.

In the wake of Pierre Bourdieu's Distinction (1979), literary studies has developed a nuanced critical apparatus for rethinking the role played by the canon and canonicity in the perpetuation of cultural capital. Our current scholarly common sense insists that far from reflecting aesthetic value, canons actually create this value socially, often thereby reinforcing hegemonic cultural values and hierarchies. And yet, even as we know this, the actual collection of texts that is consistently taught, written about, and by extension canonized remains relatively stable. By examining the frequency with which particular works are mentioned in various scholarly networks, Metacanon creates an accessible representation of this trend and in doing so introduces a greater level of transparency into the dominant allocation of literary values. In doing so, this project is similar in some respects to work being done by Mark Algee-Hewitt and Mark McGurl at the Stanford Literary Lab, although using different means and with different ends in mind. Whereas Algee-Hewitt and McGurl have produced a master list of 350 twentieth century novels by combining several "found lists" supplemented by a survey of scholars working in the field of postcolonial literature, Metacanon uses an approach that takes advantage of a wider array of harvested data drawn from thousands of journal articles. This reflects the very different goals of this project. Rather than producing a necessarily limited corpus suitable for datamining, Metacanon is primarily intended as an exhaustive but flexible representation of the canon. This allows for a much larger interactive list, but I've also found that the Metacanon list is much more diverse in terms of gender (and likely ethnicity) than the McGurl and Algee-Hewitt list, even though they consciously aimed to create a corpus that would be more representative than most standard lists. What this indicates is that although most publicly available "greatest books lists" tend to over-represent white men in their construction of literary value, scholars themselves tend to work on a much more diverse array of literary texts. In other words, there is already a working canon in existence that is much more diverse and representative than the standard lists and surprisingly more so even than Stanford's intentionally varied list; it's just that this working canon isn't generally available in an accessible, objectified form. Metacanon takes the first steps toward producing this more accessible form, even as it integrates flexibility and transparency into its framework.

The current version of Metacanon (0.6) is limited to twentieth century American fiction. As such, however, it is the most comprehensive relational database of American fiction from this period. Of course, there are more extensive listings of American literature available (for example, the Chadwyck-Healey Bibliography of American Literature). However these offer no way to easily distinguish between highly canonical works and more obscure works. In essence this forces readers looking for a definitive list of American fiction to choose between the unwieldiness of comprehensive bibliographies and the partiality of much shorter "greatest books" lists and standard field lists. What makes Metacanon unique is that it harnesses digital technology in order to offer both the expansiveness of a comprehensive bibliography while at the same time measuring the relative centrality or obscurity of each particular work.

This digital framework also allows users to become active participants in the construction of the canon rather than merely passive recipients. For example, one user might choose to see a list of the most canonical novels published between 1970 and 1979, or even more interestingly the most canonical novels of the 70s according only to data from the 80s or 90s. Another user might choose to see a list consisting only of science fiction written by women. A third user might choose to alter the algorithm to calculate canonicity based only on citations in a single journal. Contrary to the widespread fear that digital or quantitative approaches to literature are fundamentally opposed to nuance and flexibility, Metacanon demonstrates that this need not be the case, at least in so far as questions of canon formation are concerned.

While most of what is written above concerns Metacanon's value as a public humanities initiative and as an aid to students, this project has growing implications for literary scholarship more broadly. One of the most fascinating lines of inquiry in the digital humanities today is the use of quantitative textual analysis to trace the connections between literary form and reception over time. For example, scholars like Richard Jean So, Hoyt Long, Ted Underwood, and many others have used digital text mining to demonstrate relationships between particular formal features of literary texts and the social categories that govern their movement through the world, such as attributed aesthetic value and genre. As the precision of Metacanon's measurement of canonicity improves, researchers could use this data along similar lines to ascertain connections between literary form and canonicity over time.

Bibliography

Algee-Hewitt, M. and McGurl, M. (2015). Between Canon and Corpus: Six Perspectives on 20th-Century Novels. *Stanford Literary Lab Pamphlet*, **8**. http://litlab.stanford.edu/LiteraryLabPamphlet8.pdf (accessed 1 March 2016).

Long, H. and So, R. J. (2016). Literary Pattern Recognition: Modernism between Close Reading and Machine Learning. *Critical Inquiry*, **42**(2): 235-67.

Sellers, J. and Underwood, T. (2016). The Longue Durée of Literary Prestige. *Modern Language Quarterly*, **77**(3).

Mining Leitmotif in James Joyce's 'Ulysses'

Ronan Crowley
crowle01@gw.uni-passau.de
Universität Passau, Germany

Gábor Mihály Tóth
gabor.toth@maximilianeum.de
Universität Passau, Germany

James Joyce is one of the most admired, emulated and mythologised masters of twentieth-century prose. His *Ulysses* (1922) is among the cardinal texts of literary modernism produced between the world wars. The text creates its numinous effect through the repetition of short phrases over the course of a quarter of a million words. As early as 1929, critics invoked the musical device of leitmotif to explain this form of literary repetition (Curtius, 1929). Leitmotif describes a signature phrase or cue that accompanies and signals the presence of a character, locale or theme in a work. The device achieved a new importance in the nineteenth century through the opera of Richard Wagner, and it quickly migrated from music to literature in the writings of Édouard Dujardin, Thomas Mann and Marcel Proust. In Joyce's hands, the pervasive use of the device amounts to "a sort of linguistic magic-realism" (O'Callaghan, 2011). But whereas musical leitmotif is conveyed through auditory recall – listeners recognise a brief musical phrase as an instance of leitmotif – written language cannot always offer this immediacy. The problem is a singular one: how do readers and how do computational tools recognise and unify the discrete instances of leitmotif that are distributed across a text?

Methodology

The proposed paper will report on a series of experiments that combine methods of corpus and computational linguistics with close reading to gauge a fuller extent of the repetition in the novel and to assess its worth as leitmotif. From the perspective of text analysis, leitmotif is the purposeful repetition of textual fragments (or n-grams) in a text or over a collection of textual documents. Our computer assisted retrieval of leitmotifs is based on the following set of assumptions: First, for a sequence of words to be a leitmotif, it must capture a (human) reader's attention through distinctive word choice or unusual collocation. This distinctiveness ensures a reader can recall earlier occurrences. Second, a given sequence of words functions as a leitmotif only if it occurs in a limited number of chapters or episodes of *Ulysses*. The sequence must occur in more than one episode but not in all eighteen. Key to distinguishing a given sequence is a limitation in the number of its occurences. Finally, previous scholarship has observed that Joyce not only repeated leitmotifs verbatim but also paraphrased, transposed, abbreviated or otherwise altered their individual elements (see Büchler et al., 2014 in this context). A computational approach needs to take stock of this state of affairs.

Work to Date

At the current stage of our research, we focus on bigrams. By treating each episode of the novel as a single document, we have constructed a document collection to measure the inverse document frequency (idf) of bigrams in the collection. The use of idf metrics enables us to retrieve those bigrams occur in a given number of episodes. Initially we set a threshold of occurrence at two, three and four episodes; in the light of the results produced we are continually revising this threshold. As a second step, the uniqueness or distinctiveness of bigrams was investigated by measuring the probability of their occurrence in a contemporary corpus of English texts put together from documents in Project Gutenberg and Archive.org. This helped to filter out very common constructions that enjoy a high idf. Named entities were also eliminated from the list of bigrams. We have also examined bigrams for "fuzziness" – gauging whether the constituent elements reoccur but with words inserted between them or in a reversed or otherwise transposed order. We have produced a list of around two hundred bigrams that are candidates for consideration as leitmotifs. Finally, a striking feature of Joyce's style in *Ulysses* is the frequent use of compound coinages. Not only did he insert such neologisms into the text, but he also split them, reusing the constituent units of compounds in close proximity. This stylistic device also contributes to the pervasive sense of repetition in *Ulysses*, functioning as an alternative to and special case of leitmotif. To identify this type of repetition, we have retrieved and split all compound coinages in the novel, and examined whether their constituents reoccur within a given word distance.

Challenges

The main obstacle to identification is the "protean nature" of leitmotif itself (Bribitzer-Stull, 2015). Whereas a highly distinctive phrase like Joyce's "Agenbite of inwit," which occurs in *Ulysses* seven times (Joyce, 1922), can readily be identified as a leitmotif, the associative potential inherent in even very short units of language means it is not always clear how one is to distinguish leitmotif from mere linguistic flukes or from instances of more general intertextuality (Kristeva, 1967). Examples of the latter include lexical priming (Hoey, 2005) and natural language collocation – for example, if an author writes "he opened the," a reader of English would reasonably expect the next word to be "door." At the moment, our list of candidate leitmotifs is checked by close reading bigrams within the context of the sentences in which they occur in the novel. Those bigrams that are accepted or judged as valid instances of leitmotif are added to a purpose-built database.

Outcomes

In addition to extending the inventory of leitmotifs identified in *Ulysses* (see, for example, Schutte, 1982), the experiments and the subsequent analysis have another key goal. They are meant to test the hypothesis that leitmotif not only offers a way to flag the presence of a character, locale or theme in the novel, but also creates a series of "primitive hyperlinks" or "analogue hyperlinks" (Cope and Phillips, 2006) threaded throughout *Ulysses*. Critics have often invoked hypertext to explain the myriad connections established within Joyce's work (e.g. James, 1999), but our research will lead to an online hypertext edition of *Ulysses* that allows the reader to explore the non-linear reading paths created by leitmotif.

Bibliography

Bribitzer-Stull, M. (2015). *Understanding the Leitmotif: From Wagner to Hollywood Film Music*. Cambridge: Cambridge University Press.

Büchler, M., Burns, P. R., Müller, M., Franzini, E. and Franzini, G. (2014). Towards a historical text re-use detection. In Biemann, C. and Mehler, A. (eds), *Text Mining: From Ontology Learning to Automated Text Processing Applications*. Cham: Springer, pp. 221–38.

Cope, B. and Phillips, A. (2006). *The Future of the Book in the Digital Age*. Oxford: Chandos.

Curtius, E. R. (1929). *James Joyce und sein "Ulysses."* Zürich: Verlag der Neuen Schweizer Rundschau.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. New York: Routledge.

James, L. L. (1999). Some notes on Joycean hypertext: machine—tra(ns)versal— acrostic. *Litteraria Pragensia*, **9**(17): 59–89.

Joyce, J. (1922). *Ulysses*. Paris: Shakespeare and Company.

Kristeva, J. (1967). Bakhtine, le mot, le dialogue, le roman. *Critique*, **23**(239): 438–65.

O'Callaghan, K. (2011). Mapping the "call from afar": the echo of leitmotifs in James Joyce's literary landscape. In Bénéjam, V. and Bishop, J. (eds), *Making Space in the Works of James Joyce*. New York: Routledge, pp. 173–90.

Schutte, W. (1982). *Index of Recurrent Elements in James Joyce's "Ulysses."* Carbondale: Southern Illinois University Press.

WordPress as a Framework for Automated Data Capture, Filtering and Structuring Processes. The New Order of the Authors

Antonio Cruces Rodríguez
antonio.cruces@uma.es
University of Málaga, Spain

Nuria Rodríguez Ortega
nro@uma.es
University of Málaga, Spain

Carmen Tenor Polo
carmen.tenor@gmail.com
University of Málaga, Spain

The Exhibitium Project, awarded by the BBVA Foundation, is a data-driven project developed by an international consortium of research groups. One of its main objectives is to build a prototype that will serve as a base to produce a platform for the recording and exploitation of data about art-exhibitions available on the Internet. Therefore, our proposal aims to expose the methods, procedures and decision-making processes that have governed the technological implementation of this prototype, especially with regard to the reuse of WordPress (WP) as a development framework.

According to the project's purpose, it was necessary to create a device that, to the extent possible, could capture in automated way information on art exhibitions from any Internet source. Consequently, the inquiry into the possibilities of web mining strategies emerged as a priority from the early stages. Taking into account the high expressiveness and flexibility of linguistic structures usually used in the description of art exhibitions, our project opted for a mixed platform which combines the potential modeling system based on textual indicators, the heuristic means that characterize some methods -such as the Bayesian classification- and the human supervision provided by a well trained team of editors.

1. General overview. WordPress as a framework of the Expofinder system

As Baumgartner et al. (2009: 1) established, the web data extraction task is usually divided into five different functions: (1) the web interaction, which mainly comprises the navigation through predetermined web pages containing the information sought; (2) the extraction of the searched data by means of a software that identifies, extracts and transforms them into a structured format; (3) the setting of a specific calendar that enable to perform automatically the extraction tasks in regular sequence; (4) the processing of the captured data, which includes filtering, transformation -if applicable-, refining and integration; and (5) the delivery of the resulting structured data to a variety of data analysis-based systems.

Assuming this distribution as the most convenient for our purposes, we decided to include them in the Exhibitium's architecture grouped into two large blocks.

A. A block consisting of an automated capture system of information robust enough to ensure the reliability of the collected data.

B. A second block made up by the set of elements necessary to store the data, including functions for filtering, cleaning, management, structuring and description. This block also incorporates a system to export the collected data to those platforms that will process and analyze them during the second phase of the project.

Block A was called Beagle, and block B became known as ExpoFinder. Both blocks work in a coordinated manner, so that what is extracted by Beagle is put at the disposal of ExpoFinder. The two blocks are part of an unified system configured by a cyclic algorithm: Beagle captures, ExpoFinder analyzes and approves the captured information, the team of editors validates or discards what ExpoFinder has previously approved, and Beagle recaptures again (see figure 1).

Regarding the software, after preliminary versions based on own developments, it was decided that the most interesting option between the free software and open source solutions currently available (as the openness philosophy is a sine qua non requirement of this project) would be to use WP as framework of the system. The main benefits that the use of WP as framework offers for our project can be synthesized in the following items: a database with a flexible and very solid organizational structure; a layer of a core application with numerous hooks which allow to maximize its functionality; and a high easy management system to carry out tasks on the two sides (server and client), assuming both administrator and user roles. Thus, for the implementation of the Beagle-ExpoFinder system we took advantages of the predefined data base, the available APIs and the set of data visualization templates to build solutions using an application that is already fully functional.

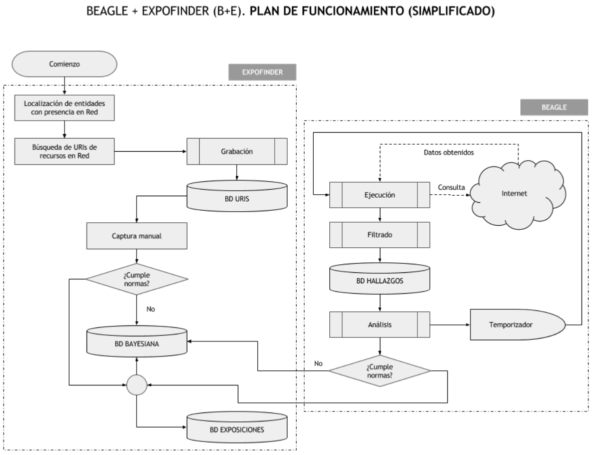
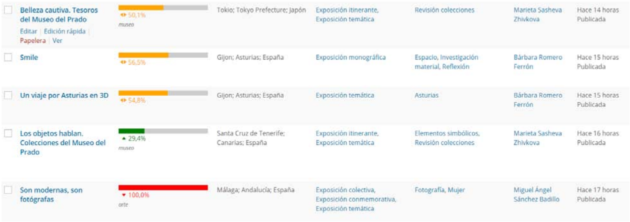


Figure1. Beagle + ExpoFinder. Operating plan (simplified)

We used WP without adaptations, that is, as it can be downloaded from the Internet. All the functionality of our application lies, then, on the code itself that constitutes WP, so it is not supported on variant versions (forks) of the original program. Hence, any improvement provided by the computing community will be directly usable by our project without further adaptations. As part of the requirements of the development of the Beagle-ExpoFinder system (B + E), from the beginning it was considered that the programming work did not constitute a «tailored suit» for the Exhibitium project. On the contrary, we expect that this work can be useful in other projects with a small number of modifications or by using configuration files or other similar systems. For that reason, our choice was to implement B + E by means of a «WP theme», solution that easily allows us to readapt the software to different purposes.

2. Beagle and Expofinder development and technical features

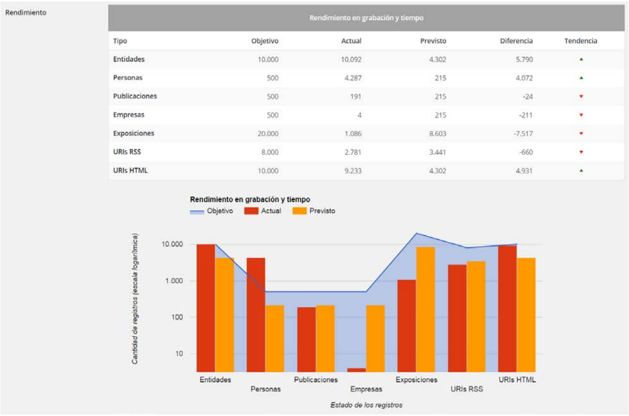
Beagle captures – as it has been said- by automated means web data concerning temporary art exhibitions from any source of information, and includes a filtering mechanism. The automated capture process uses the tools of WP API, particularly WPCron. Likewise, the frequency of the process is configured according with the options offered by ExpoFinder to the system administrator. Beagle employs two statistical complementary functions to «predict» the degree of the adequacy of the captured information to the ExpoFinder preconditions: 1. One of this is based on the intersection of a set of «positive» and «negative» keywords with a proportional weight assigned to each one, which is also based on the shortest path algorithm of Bellman-Ford; 2. The other is defined by its heuristic nature; it employs a naive Bayesian classifier to guide the «human» editor during the task of discriminating whether or not an information captured by Beagle is valid. The latter is able to improve their efficiency through continuous learning processes (each discarding or acceptance made by the «human» editor refines the system «perceptiveness» (see figures 2 and 3).



□	Belleza cautiva. Tesoros del Museo del Prado Editar | Edición rápida | Papelera | Ver	50,1% museo	Tokio; Tokyo Prefecture; Japón	Exposición itinerante, Exposición temática	Revisión colecciones	Marieta Sasheva Zhivkova	Hace 14 horas Publicada
□	Smile	56,5%	Gijon; Asturias; España	Exposición monográfica	Espacio, Investigación material, Reflexión	Bárbara Romero Ferrón	Hace 15 horas Publicada
□	Un viaje por Asturias en 3D	54,8%	Gijon; Asturias; España	Exposición temática	Asturias	Bárbara Romero Ferrón	Hace 15 horas Publicada
□	Los objetos hablan. Colecciones del Museo del Prado	▲ 29,4% museo	Santa Cruz de Tenerife; Canarias; España	Exposición itinerante, Exposición temática	Elementos simbólicos, Revisión colecciones	Marieta Sasheva Zhivkova	Hace 16 horas Publicada
□	Son modernas, son fotógrafas	▼ 100,0% arte	Málaga; Andalucía; España	Exposición colectiva, Exposición conmemorativa, Exposición temática	Fotografía, Mujer	Miguel Ángel Sánchez Badillo	Hace 17 horas Publicada

Figure 2. Screenshot. Exhibitions list (fragment). See the Bayesian classifier indicator

ExpoFinder also includes a control system (QC) that identifies the mistakes and failures, which are also associated with the human editor who made them, so that he/she can perform the appropriate corrections (see figures 3 and 4).



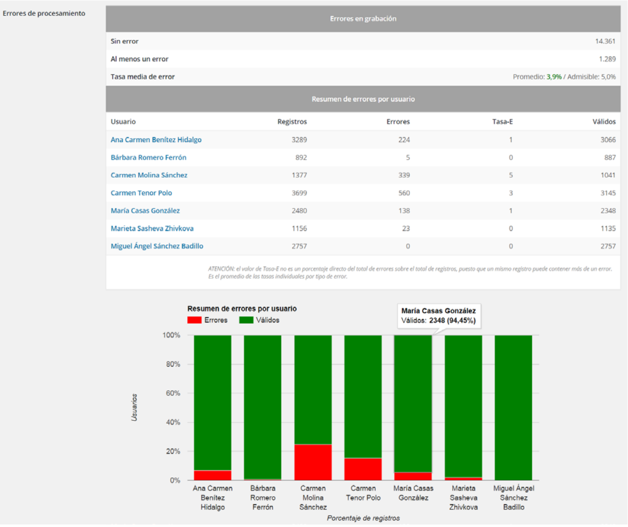
Rendimiento

Rendimiento en grabación y tiempo

Tipo	Objetivo	Actual	Previsto	Diferencia	Tendencia
Entidades	10.000	10.092	4.302	5.790	▲
Personas	500	4.287	215	4.072	▲
Publicaciones	500	191	215	-24	▼
Empresas	500	4	215	-211	▼
Exposiciones	20.000	1.086	8.603	-7.517	▼
URIs RSS	8.000	2.781	3.441	-660	▼
URIs HTML	10.000	9.233	4.302	4.931	▲

Figure 3. Screenshot. Automated evaluation of efficiency

In its current state of development, the Beagle-ExpoFinder system captures and selects daily about 100 references from more than 12,000 web sources of information. Its error rate during the recording and validation processes is about 3.9%, below the 5% initially considered as permissible.



Errores de procesamiento

Errores en grabación	
Sin error	14.361
Al menos un error	1.289
Tasa media de error	Promedio: 3,9% / Admisible: 5,0%

Resumen de errores por usuario

Usuario	Registros	Errores	Tasa-E	Válidos
Ana Carmen Benítez Hidalgo	3289	224	1	3066
Bárbara Romero Ferrón	892	5	0	887
Carmen Molina Sánchez	1377	339	5	1041
Carmen Tenor Polo	3699	560	3	3145
María Casas González	2480	138	1	2348
Marieta Sasheva Zhivkova	1156	23	0	1135
Miguel Ángel Sánchez Badillo	2757	0	0	2757

ATENCIÓN: el valor de Tasa-E no es un porcentaje directo del total de errores sobre el total de registros, puesto que un mismo registro puede contener más de un error. Es el promedio de las tasas individuales por tipo de error.

Figure 4. Screenshot. Quality control (QC). Resume

Bibliography

Baumgartner, R. et al. (2009). Web Data Extraction System. In *Encyclopedia of Database Systems*. Springer-Verlag.

Kokkoras, F. et al. (2013). DEiXTo: A Web Data Extraction Suite. *Proceedings of 6th Balkan Conference in Informatics* (BCI 2013), Nueva York: ACM, pp. 9-12.

Pree, W. (1994). *Design Patterns for Object-Oriented Software Development*. Reading, Massachussets, USA: Addison-Wesley, ACM Press Books.

Raposo, J. et al. (2002). The Wargo System: Semi-Automatic Wrapper Generation in Presence of Complex Data Access Modes. *Database and Expert Systems Applications. Proceedings*, 13th International Workshop, IEEE, pp. 313-17.

Notes

[1] Generation of knowledge about temporary art exhibitions for a multivalent reuse was the topic of the proposal presented to the 2014 competition organized by the BBVA Foundation for projects in the field of Digital Humanities, resulting selected from over 250 submissions. The project website is available at: http://exhibitium.com. The Exhibitium Project began in January 2015 and will end in December 2016, so currently we are completing the first phase.

[2] They are: iArtHis_Lab (http://www.iarthislab.es) and Khaos (http://khaos.uma.es) at the University of Málaga; Techne, ingeniería del conocimiento y del producto (http://www.ugr.es/~tep028/quienes_somos_es.php) at the University of Granada; and CulturePlex at the University of Western Ontario (http://www.cultureplex.ca).

[3] Specifically, the ultimate Exhibition's purpose is to extract unprecedented and strategic knowledge about temporary art exhibitions through the use of a variety of data mining techniques.

[4] Although, in reality, according to the Tom McFarlin's statement in his popular page «tuts +» (http://tutsplus.com/), it is more a foundation that a framework. And maybe he is right: a framework consists of a set of conventions as well as libraries and tools that allow us to easily start working on an application. In short, it provides the means by which an application can be built from scratch, from the database schema to the front end. However, a foundation allows to «extend» an existing application. WP has its own well-defined internal mechanisms, and the foundation simply expands its operation or takes advantage of it for their own benefit.

[5] The advantages that a robust mechanism as such provided by WP offers for the maintenance of a security system (essential in any development accessible through Internet), or the substantial savings in time and resources involved in a CRUD structure records management -which is both sufficiently malleable to suit any need and rigid enough to follow canonical deployment patterns (such as the «nonce» safeguards in the capture forms), are weighty arguments when opting for the use of one framework or another.

[6] Even though this document is not largest enough to expose in detail the list of selected preconditions of significant terms used by Beagle in order to filter the captured information, we want to emphasize that this is a «weighted» relationship of lexemes in which each root term is assigned a total «weight» in the set (1 to 3). When the entire process is complete, the absolute amount of the sum corresponding to the found terms is weighted with the relative values (relating to the length of the text where they have been detected) to assign a positive or negative evaluation to the whole information.

[7] The Bellman-Ford algorithm (or Bellman-Ford-Moore) calculates the shortest paths from a single source vertex to all other vertices in a weighted digraph. It is slower than Dijkstra's algorithm for the same problem, but more versatile, since it is suitable to deal with graphs using negative numbers for edge weights. ExpoFinder takes advantage of it in its weighting mechanism, useful for us because we work with lists of lexemes for words used as «positive» or «negative» markers.

[8] In machine learning terminology, the «naive Bayesian» constitutes a family of simple probabilistic classifiers based on the application of Bayes' theorem about the hypothesis of independence between variables. Widely studied since the 1950s, it began to be used since the beginning of the next decade as a taxonomy method capable of self-optimization in the recovery community text. We use the frequency of occurrence of a given lexeme as a trigger, so that ExpoFinder can contribute to the semi-automated selection of relevant information from the experience gained. It is not a pure discriminative mechanism, but an auxiliary tool that has proven to be useful for application operators.

The Project Zeri Photo Archive: Towards a Model for Defining Authoritative Authorship Attributions

Marilena Daquino
marilena.daquino2@unibo.it
University of Bologna, Italy

Silvio Peroni
silvio.peroni@unibo.it
University of Bologna, Italy

Francesca Tomasi
francesca.tomasi@unibo.it
University of Bologna, Italy

Fabio Vitali
fabio.vitali@unibo.it
University of Bologna, Italy

Introduction

The global project of data conversion of the notable Italian 'Zeri Photo Archive' into a Linked and Open Dataset[1] primarily regarded the analysis of the **description model of the available records**, so as to define a collection of suitable ontologies to describe such a complex domain.

Indeed, the uniqueness of the Zeri collection[2], which includes about 290,000 photographs of works of art and monuments, lies in the rich documentation of the described artefacts, mostly related to provenance, attributions, restoration events and their connections to the collection of 46,000 books and 37,000 auction catalogues (Mambelli, 2014).

The full project entails, together with the ontological modeling, the production of the RDF dataset, the creation of proper links to the LOD cloud, and the definition of the user interface for browsing the dataset.

F Entry Ontology

The first activity of the project had been the **formalization of the** Scheda F ('Fotografia') (Mibact, 1999) , – the metadata standard of the Italian Istituto Centrale per il Catalogo e la Documentazione (ICCD) – by mapping the schema onto CIDOC-CRM model (Crofts et al., 2011). In the conversion we initially considered the specific flavor of the Scheda F used by the Zeri Foundation, i.e. its subset of 113 fields (based on the experimental 1.04 version of the ICCD standard) and an handful of custom extensions to it. A deep analysis of the schema of the Scheda F showed that it is organized in semantically independent sections (called "paragrafi", or *paragraphs*), each one belonging to a specific FRBR concept (Work, Manifestation, Item); this allows the mapping to proceed by logical sections affecting only a limited number of entities and relating these entities to the data documented by the fields of the schema.

Before accomplishing the mapping, we proceeded with the creation of a **new ontology called FEO** (*F Entry Ontology*[3]). Since our final goal was to make available Scheda F data in a triple store, the target language we chose was OWL 2 DL. The current version of FEO introduces the classes and properties that characterize three specific concepts: the photograph, the work of art that is the subject of the photograph, and the Scheda F itself describing the photograph and its subject.

So, through the use of well-known ontologies – i.e. CIDOC-CRM, but also PROV-O (Lebo et al., 2013), and FaBiO (Peroni and Shotton, 2012), as part of the SPAR Ontologies[4] (Peroni, 2014) – plus the FEO ontology developed *ad hoc*, most of the content expressed as descriptive entries in the Scheda F have been already formally represented (Gonano et al., 2014).

OA Entry Ontology

In this presentation we propose an extension of our work on the Zeri Photo Archive by introducing a **new ontology** for representing works of art and their related information, namely, the OA Entry Ontology[5]. In particular, this ontology is based on the *Scheda OA* ('Opera d'Arte') – another ICCD metadata standard[6] – and proposes a mapping between the content standard and, again, the CIDOC-CRM, in order to create shareable descriptions of metadata[7]. In addition, other kinds of information that are not easily representable through the aforementioned standards, such as certain peculiar relations between works of art, are modelled by means of new classes and properties created in the OA Entry Ontology. This last allows to describe, in particular, the work of art and the related items, by focusing on some classes (e.g. copy, derivation, fake, drawing, model, replica, sinopia) and by using properties as necessary connection typization (e.g. conceived or former).

HICO and authorship attribution

Moreover, in this paper we further investigate a way for providing a clear and shareable representation of questionable information, i.e., the **authorship attribution of works of art**. In the Zeri Photo Archive there are particular authorship attribution created by either the Zeri Foundation cataloguers and/or by Federico Zeri (collector of photographs) himself, and such attributions (that could be accepted or discarded at certain point) are accompanied with the criteria that corroborate the cataloguers' choice.

In order to provide a precise characterisation of all these aspects, we discuss the adoption of **HiCO**, Historical Context Ontology (Daquino and Tomasi, 2015) as a way for enabling a definition of authoritative attributions based on Zeri cataloguers' own criteria. HiCO[8] is an OWL 2 DL ontology we created for describing contents of data (e.g., an authorship attribution), in particular cultural heritage data, and data creation itself (e.g., RDF statements representing above mentioned authorship attribution) as part of an interpretative process. Cultural heritage object is a wide concept: it includes any sort of representation of culture heritage embodied in a tangible form like artifacts (books, documents, and, as in this case, works of art), but also any concept, assertion and interpretation somehow bounded to cultural objects.

With the hico:InterpretationAct class it's possible to **describe the interpretation act as a process**:

- the conceptualization of the interpretation and its classification, for enabling further relations among different kind of interpretations;
- the embodiment of the interpretation as RDF statements, for representing information extracted from the content of the object of interest.

Two fundamental object properties complete the process: the hico:hasInterpretationType property and the hico:hasInterpretationCriterion property. The former states an arbitrary classification of the interpretation, which can be simply defined as philological, historical, semiotic, linguistic etc. The latter is a briefly explanation of the criterion used to state information extracted from a source, e.g. a literally transcription, a hypothesis, or the adoption of the literature about a specific argument.

A crucial aspect of the project is the **correct formalization of statements** so as to allow the ontologically-consistent coexistence of data created by different actors that express contradictory statements about the same subject (e.g., authorship attribution data of a work of art), in order to guarantee the data integration with Pharos (Reist et al., 2015) project partners, of which the Zeri Foundation is a member. By the use of SWRL rules applied to relations between sources, criteria and context information used by an agent to explain his interpretation, **we could formally infer when an interpretation can be considered authoritative**.

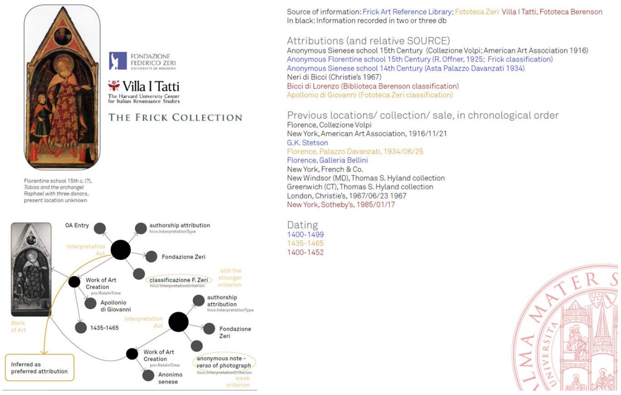
An example of authoritative assertion

We could give an **example**. We state aside the interpretation (i.e. the assertion "An artist X is author of a specific work of art Y in a specific time Z") which sort of interpretation we are dealing with (i.e. authorship attribution), and which criteria have been used by the cataloguer to assert such proposition. A provenance statement ensures both author of assertion (i.e. interpretation) and author of data conversion are fully described, ensuring that no contradictory statements will affect data validity.

When necessary conditions for stating authorship are fulfilled, an attribution may be inferred as authoritative. In the example (Fig. 1), we have an attribution which respects minimal requirements for being considered an **authoritative assertion**:

- it has been stated by a well-known author (i.e. the cataloguer of Zeri Foundation);
- it considers as criterium an authoritative source (i.e. the photo depicting the work of art);
- it agrees with another interpretation, i.e. Federico Zeri classification.

This obviously doesn't entail that the attribution is surely correct, but it can be a useful tool for historians when searching for different attributions and related criteria adopted in interpretative process.


FONDAZIONE FEDERICO ZERI UNIVERSITÀ DI BOLOGNA
Villa I Tatti The Harvard University Center for Italian Renaissance Studies
THE FRICK COLLECTION
Florentine school 15th c. (?), Tobias and the archangel Raphael with three donors, present location unknown
OA Entry
authorship attribution
hico:InterpretationType
Interpretation Act
Fondazione Zeri
Work of Art Creation
pro:RoleInTime
classificazione F. Zeri
hico:InterpretationCriterion
still the stronger criterion
Apollonio di Giovanni
1435-1465
Work of Art
authorship attribution
hico:InterpretationType
Interpretation Act
Fondazione Zeri
Work of Art Creation
pro:RoleInTime
anonymous note - verso of photograph
hico:InterpretationCriterion
weak criterion
Anonimo senese
Inferred as preferred attribution

Source of information: Frick Art Reference Library; Fototeca Zeri Villa I Tatti, Fototeca Berenson
In black: Information recorded in two or three db

Attributions (and relative SOURCE)
Anonymous Sienese school 15th Century (Collezione Volpi; American Art Association 1916)
Anonymous Florentine school 15th Century (R. Offner, 1925; Frick classification)
Anonymous Sienese school 14th Century (Asta Palazzo Davanzati 1934)
Neri di Bicci (Christie's 1967)
Bicci di Lorenzo (Biblioteca Berenson classification)
Apollonio di Giovanni (Fototeca Zeri classification)

Previous locations/ collection/ sale, in chronological order
Florence, Collezione Volpi
New York, American Art Association, 1916/11/21
G.K. Stetson
Florence, Palazzo Davanzati, 1934/06/25
Florence, Galleria Bellini
New York, French & Co.
New Windsor (MD), Thomas S. Hyland collection
Greenwich (CT), Thomas S. Hyland collection
London, Christie's, 1967/06/23 1967
New York, Sotheby's, 1985/01/17

Dating
1400-1499
1435-1465
1400-1452

Fig. 1 Sample of multiple interpretation of the same object and Zeri authoritativeness

Conclusions

To conclude, as said before it will be the natural completion of the Zeri Project the RDF triple store set up, the creation of links to other datasets, and the definition of the user interface for browsing the linked open dataset. All these activities are now object of our research group industry and the final publication of the project is expected for the middle of 2016. Data had already been transformed in a RDF/XML dataset, according to above mentioned ontologies. Next steps of the project involve thus data publication, ensuring access to them through a SPARQL endpoint. When published, data will be enriched with further RDF links to major datasets and authority files online (e.g. VIAF for people and works, Getty thesauri, GeoNames for places[9]). This ensures our data will really be part of the LOD Cloud, avoiding creation of another data silo. So enriched data will then be exploited in a new smart application, which will enable users to search data about both photos and works of art. Through this modelization data and new relations will be easier discovered, enhancing user experience.

Bibliography

MiBACT (1999). *Strutturazione dei dati delle schede di catalogo: beni artistici e storici: scheda F, prima parte*. Roma: ICCD.

Mambelli, F. (2014). Una risorsa online per la storia dell'arte:

il database della fototeca Zeri. In: Ciotti, F. (Ed.), *Digital Humanities: progetti italiani ed esperienze di convergenza multidisciplinare*. Quaderni Digilab, Università di Roma La Sapienza.

Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M. (2011). *Definition of the CIDOC Conceptual Reference Model* (version 5.0.4). http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf.

Lebo, T., Sahoo, S. and McGuinnes, D. (2013). *PROV-O: the PROV Ontology*. W3C Recommendation. http://www.w3.org/TR/prov-o/.

Peroni, S. and Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, **17**: 33-43.

Peroni, S. (2014). The Semantic Publishing and Referencing Ontologies. In: Peroni, S., *Semantic Web Technologies and Legal Scholarly Publishing*. Cham, Switzerland: Springer, pp. 121-93.

Gonano, C.M., Mambelli, F., Peroni, S., Tomasi, F. and Vitali, F. (2014). Zeri e LODE. Extracting the Zeri photo archive to Linked Open Data: formalizing the conceptual model. *Proceedings of the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL 2014)*. IEEE, pp 289-98.

Daquino, M. and Tomasi, F. (2015). Historical Context (HiCo): a conceptual model for describing context information of cultural heritage objects. *Communication in Computer and Information Science*, Berlin: Springer Verlag, pp. 424-36.

Reist, I., Farneth, D., Stein, R. and Weda, R. (2015). An Introduction to PHAROS: Aggregating Free Access to 31 Million Digitized Images and Counting. Speech at *CIDOC 2015*. New Dehli.

Notes

[1] The project is supported by the Zeri Foundation with the University of Bologna and started in 2014

[2] Fondazione Zeri, Photo Archive Catalog, http://catalogo.fondazionezeri.unibo.it

[3] http://www.essepuntato.it/2014/03/fentry

[4] http://www.sparontologies.net

[5] http://purl.org/emmedi/oaentry

[6] See the ICCD cataloging standards at: http://www.iccd.beniculturali.it/index.php?it/473/standard-catalografici

[7] We are planning to publish both the F and the OA mapping to CIDOC-CRM in the next few months, according with the ICCD activities

[8] http://purl.org/emmedi/hico

[9] A first check on https://datahub.io/dataset

Materiality and Metadata of Digitised Photographs: A Theoretical Inquiry

Vinayak Das Gupta
vinayak.dasgupta@gmail.com
Maynooth University, Ireland

This paper suggests a separation between the content, the carrier, and the presentational forms of the digitised photograph (as opposed to a born-digital image) and presents the draft for a metadata schema that is able to record the circulation of the object. The arguments presented in this paper are adapted from the research conducted for a recently-submitted doctoral thesis.[1]

Scholarly study of the photographic image, specifically with the 'material turn' in anthropology and cultural studies,[2] has attempted to inspect the photograph as a physical object (Batchen, 1997, Edwards, 2002, Edwards and Hart, 2004). This process of inspection expresses the complexity of the social existence of objects and allows the investigation of the photograph as a material object whose understanding is augmented by its form. Photographs have specific modes of circulation, production, and consumption, and their inspection has potential beyond the critiques of representation alone.[3] The inherent bias of the indexical qualities of the image over its material properties may overlook the social and cultural contexts within which the photograph is born and used. While current digitisation techniques have found accurate methods of copying the content image, the description of its materiality remains a challenge. If the physicality of the photograph is central to its understanding, this paper inspects the possibilities of providing the digital referent with the material information. It presents an examination of both the materiality of the photographic image (and its transformation into a digital object), and the means through which the presentational forms of the original may be inscribed in the digital referent.

The digitised photograph is a material object: to observe this materiality, a separation between the content and the carrier of the object is required. The photographic image is printed on paper, and this paper is the carrier of the content image. Similarly, the content of the book is carried in the physical form of the book — the paper, glue, and ink that hold it together. In the physical object (as opposed to its digital counterpart) the content and the carrier are closely inter-linked to the point where their separation is difficult. However, the carrier may change at different moments of time, which may provide the object with different contexts: consider an image that was first printed on photographic paper, then printed in a newspaper, and then, perhaps, in a book. The different material existences of the image provide contexts that locate it within different points in history. The digital object is, similarly, carried by electronic circuits. That the digital object has materiality, then, is undeniable. The problem, perhaps, in identifying this materiality lies in the degrees of separation between the circuits and the perception of the object. To view an electronic image, a screen — an enabler — is required. The experience of the object then is governed by an intermediary system. The dislocation between the carrier of the object and the experience of the object is perhaps the source of a material fallacy.[4] If it is acceptable to think of this separation between the content and the carrier, it is possible to argue that the digital object is merely one iteration of a different carrier. The digitised image, then, is the original image, in a new material form.

The photograph has multiple lives; it exists within sociocultural contexts, and to understand it, the content image must be seen in conjunction with its material form. Since the inception of photography, photographic images have been used in a variety of contexts, and have been presented in a multitude of ways; the carrier often determines the use the content image is put to. Whether preserved in photoalbums (arranged thematically or sequentially), sold as postcards for the curious, or published as documentary evidence the presentational form of the photograph weighs heavy on the readings of the image. Presentational forms, in particular, guide the way in which photographs were used after their inception, and also the way they were understood. It is, here, important to distinguish between the carrier and the presentational form: the carrier is always material, while the presentational form is ideological. The materiality guides the technical production of the image, and bears the imprint of time on it. The presentational form reveals the social, political, cultural, and religious contexts within which the photograph is *used*.[5]

This paper proposes to include the material specifics of the physical photograph at the level of the metadata of the digitised image. Metadata schemata for visual resources, such as CDWA and VRA Core, articulate the description of objects by distinguishing between Work and Image.[6] This separation attempts to describe the complex relationship between the original and the surrogate. In a similar vein, this paper suggests a separation between the content, the carrier, and the presentational form of the photograph in its descriptive schema. For photographic images, it is more important to record its material transformations than separate the original from the copy.[7] While the content image remains the same, different material specifics of the photograph change as the object adopts new carriers and new presentational forms. The proposed schema is able to record multiple carriers and presentational forms for the same photographic object. This would, potentially, help to examine the circulation of the photographic object — a key concern of scholarly research in the field. This schema also presents the possibility of being extended to a linked data format that multiple curators can add to in order to articulate information about the same object. The paper presents a blueprint for the proposed schema — the structural and the functional aspects of its design.

Bibliography

Batchen, G. (1997). *Photography's Objects.* Albuquerque: University of New Mexico Art Museum.

Benjamin, W. (1969). The Work of Art in the Age of Mechanical Reproduction. In Arendt, H. (ed.), Zohn, H. (tr.), *Illuminations: Essays and Reflections,* New York: Schocken, pp. 224.

Das Gupta, V. (2015). From Evidence to Essence: Curating Thematic Collections with Photographs from the British Raj. *Ph.D. thesis*, Trinity College Dublin.

Edwards, E. (2002). Material beings: objecthood and ethnographic photographs. *Visual Studies*, **17**(1): 67-75.

Edwards, E. and Hart, J. (2004). Photographs as Objects. In Edwards, E. and Hart, J. (eds.), *Photographs Objects Histories*, London and New York: Routledge.

Gabler, H. W. (2007). The Primacy of the Document in Editing. *Ecdotica* **4**: 197-207.

Miller, D. (1987). *Material Culture and Mass Consumption*. Oxford: Basil Blackwell, pp. 9.

Sassoon, J. (2004). Photographic materiality in the age of digital reproduction. In Edwards E. and Hart, J. (eds.), *Photographs Objects Histories*, London and New York: Routledge.

Tagg, J. (1988). *The Burden of Representation: Essays on Photographies and Histories.* Minneapolis: University of Minnesota Press.

Notes

[1] Thesis titled 'From Evidence to Essence: Curating Thematic Collections with Photographs from the British Raj' submitted in partial fulfilment of the requirements for the degree of Ph.D to Trinity College Dublin on 30th September 2015.

[2] Material cultural analysis, from an anthropological position, questions the assumed superiority of language over other forms of expression (Miller, 1987).

[3] John Tagg (1988) discusses the manner in which the currency and the value of material objects arise in certain distinct and historically specific social practices.

[4] Joanna Sassoon (2004) presents an engaging examination of the materiality of digital photographs. However, the central rationale in her paper takes premise on the assumption that digital objects have no material existence.

[5] A similar concern may be seen in the area of textual editing with concerns about the primacy of the document over the text. For further reading see, for instance, Hans Walter Gabler's 'The Primacy of the Document in Editing' (2007).

[6] VRA Core 4, in fact, is built around three record types — Work, Image, and Collection.

[7] Walter Benjamin in his seminal essay 'The Work of Art in the Age of Mechanical Reproduction' contends that for photographs, asking for the mechanically produced, original print is non-sensical.

Notes from the Transcription Desk: Modes of engagement between the community and the resource of the Letters of 1916

Vinayak Das Gupta
vinayak.dasgupta@gmail.com
Maynooth University, Ireland

Neale Rooney
neale.rooney@gmail.com
Maynooth University, Ireland

Susan Schreibman
susan.schreibman@gmail.com
Maynooth University, Ireland

The *Letters of 1916* is Ireland's first public humanities project. It collects, digitises, transcribes, encodes, and makes available through its electronic platform epistolary documents written between 1st November 1915 and 31st October 1916. The year 1916 was one of transition for Ireland: between its involvement in the Great War and the rise of militant nationalism, the country was divided by sentiment, separated by ideals. 2016 sees the centennial commemoration of the Easter Rising[1] across Ireland. A more complex interpretation of the events that transpired in Easter week 1916 has entered national consciousness and with it the interest in the smaller and more personal accounts of those caught up in the ensuing violence. The rhetoric of the letter presents personal perspectives and individual memory traces; the collected letters provide an insight into those fragmentary stories that, inspected together, constitute a collective consciousness. Letters are shared experiences that connect people across geographical spaces.[2] The unique personal perspective of the epistolary form challenges the perceptions of established history, questioning the role of memory and the acts of commemoration that this era suggests.

Since Jeff Howe's coining of the term 'crowdsourcing' in 2006 (Howe 2006), a number of pioneering projects have provided legitimacy and validity to the process.[3] The *Letters of 1916* considers crowdsourcing in the widest possible sense of the term; the processes of collection, transcription, and curation are done through public engagement. The focus of this paper rests on the volunteer community associated with the project[4] and provides an inspection of the levels of its engagement, a study of its interests and motivations, and how future projects can adapt this investigation into community interaction. Sharon Leon, discussing her observations while working on the *Papers of the War Department 1784-1800*, identifies and categorises these motivations into six fields: (a) an interest in history, (b) a sense of civic duty, (c) a specific point of scholarly research, (d) engagement based on genealogical and family research questions, (e) educational assignments, and (f) curiosity about how the transcription tool and process works (Leon 2014). This paper extends these assumptions, inspecting specifically the affect of the epistolary form on the transcriber. The topical nature of this project raises questions beyond the ones that Leon raises.

The investigation in this paper traces the manner in which the community engages with the content, a hundred years since they were written, not merely as historical documents but as individual memory traces that express personal sentiments. The centenary creates a renewed vigour in the study of these documents and the paper questions if the engagement with these letters produce a more nuanced understanding of this conflicted time. As Leon points out, one of the driving forces of community transcription lies in an active interest in scholarly research. This paper attempts to understand this very engagement in the *Letters of 1916* project; some of the letters in the resource, particularly those received from personal collections, have been unavailable to the public until this point. Does the possibility of discovering these little stories and personal narratives that are weaved within the politics of the time, create an interest that emphasises the novelty of discovery? As the transcriber actively engages and researches these documents, do her motivations lie in the possibility of unearthing new knowledge?

A data-driven examination of transcriber-activity, as evidenced in the project, suggests that individual members of the community create self-fashioned roles. The transcriber who reads and re-authors these letters, the encoder who attributes TEI[5] markup to the transcribed text, and the researcher who provides contextual information for the letter are all employed in the production of these electronic documents but their engagement is at different levels. This paper attempts to understand this division of roles based on the modes of engagement that are apparent. Re-authoring of these letters raises another question: as Barthes suggests, the relationship between the reader and the writer is complex and deeply problematic (1977). Does the re-authoring of these letters create a deeper investment in the narrative? This investment is reiterated when we consider the private emails exchanged between members of the *Letters of 1916* volunteer network and the project staff where concerns are raised over the urgency, validity, and authority of the initial transcription; the accuracy and the model of the transcription becomes a point of contention between the members of the community.

This paper considers the proposed questions in three stages and at three levels to gain a clearer understanding of the role of the transcriber within the *Letters of 1916* project. In the first instance, a statistical inspection of the metadata for the transcribed letters provides an examination of the volume and the rate of transcription over the course of the project. The engagement of the individual transcriber with specific themes[6] within the collection is revealed in the process; this mode of investigation aids in viewing the community, not as a homogenous cluster, but as individuals with specific interests and different points of engagement. The visualizations generated from these analyses illustrate the interests of the community at a macro level. In the second instance, the *Letters of 1916* volunteer network is approached with a focused survey. This ongoing survey presents the second phase of coordinated feedback that this project records. The questionnaire is designed to cover a range of topics that both reinforce the results of the statistical analysis and ask questions that lie outside the scope of data-driven inquiry. The focused design of the questionnaire is informed by the preliminary phase and attempts to tease out the motivations that lie at the heart of crowd-sourced projects such as the *Letters of 1916*. In the third and final instance, the most-prolific transcribers[7] in the project are approached for interviews regarding their role in the process of transcription. These interviews are detailed and provide a closer study of the desires of the community. The methods utilised in this paper move from a macro to a micro level, from the data to the individual, in order to derive a concrete and axiomatic base to study the modes of engagement that the *Letters of 1916* project and, perhaps, all crowdsourced projects have.

The success of a crowdsourced project lies in creating effective engagement between the community and the resource. This paper provides an investigation of the motivations and the desires of individuals that drive these forms of public history projects forward. The topical nature of this research, combined with the affect of epistolary documents, creates a unique opportunity for this study, a model for future projects to further develop their volunteer community, and the critical foundations on which future study of community transcriptions may be based.

Bibliography

Altman, J. G. (1982). *Epistolarity: Approaches to a Form.* Columbus: Ohio State University Press.

Barthes, R. (1977). Death of the Author. In Heath, S (tr.), *Image, Music, Text,* New York: Hill and Wang.

Caulfield, M. (1995). *The Easter Rebellion: The outstanding narrative history of the 1916 Rising in Ireland.* Dublin: Gill & McMillan Ltd.

Howe, J. (2006). *The Rise of Crowdsourcing.* WIRED. Available at: http://www.wired.com/2006/06/crowds/ [Accessed on 27 Oct. 2015].

Leon, S. M. (2014) Build Analyse and Generalise: Community Transcription of the *Papers of the War Department* and the Development of Scripto. In Ridge, M. (ed.), *Crowdsourcing Our Heritage*. Dorchester: Ashgate.

Papers of the War Department. (2012). Wardepartmentpapers.org. Available at: http://wardepartmentpapers.org/ [Accessed 27 Oct. 2015].

TEI: Text Encoding Initiative (2013). tei-c.org. Available at: http://tei-c.org/index.xml/ [Accessed on 27 Oct. 2015].

UCL Transcribe Bentham. (2016). Blogs.ucl.ac.uk. Available at: http://blogs.ucl.ac.uk/transcribe-bentham/ [Accessed 27 Oct. 2015].

What's on the menu? (2016). Menus.nypl.org. Available at: http://menus.nypl.org/ [Accessed 27 Oct. 2015].

Notes

[1] The Easter Rising was an armed insurrection launched by a minority of Irish nationalists against the British Empire in 1916. The Rising was contained to Dublin and suppressed following a week of fighting. It is seen as the genesis of the later Irish War of Independence. For further reading see, for instance, Max Caulfield's *The Easter Rebellion* (1998).

[2] Janet Gurkin Altman, contemplatig about the form of the letter, contends that while letters connect two geographical points they also serve as a bridge between the sender and the receiver. The epistolary author can either choose to emphasisne either the bridge or the distance (1982).

[3] Since Howe's definition of the term there have been several pioneering project based on crowdsourcing. For instance, see *Transcribing Bentham*, *Papers of the War Department 1874-1800* and *What's on the Menu*?

[4] As of October 2015 there are 1159 registered users on the *Letters of 1916* site. These users transcribe, on average, 192,409 characters a month.

[5] The Text Encoding Initiative (TEI) is a consortium which develops and maintains a standard for the representation of texts in their electronic forms. For further information see http://www.tei-c.org/index.xml

[6] Letters can belong to one or multiple themes including *The Easter Rising*, *World War I*, *Family Life*, *Love Letters*, *Official Documents*, *Politics*, the *Irish Question* and more.

[7] Each month the *Letters of 1916* Project generates a list of "top" transcribers. This list is determined by the number of characters a specific individual has transcribed in that time frame.

Automatisation Du Workflow Audiovisuel, Quel Impact Sur Le Spectateur?

Charles-Alexandre Delestage
charles-alexandre.delestage@univ-valenciennes.fr
Université de Valenciennes et du Hainaut-Cambrésis, France

Sylvie Leleu-Merviel
sylvie.merviel@univ-valenciennes.fr
Université de Valenciennes et du Hainaut-Cambrésis, France

Alain Lamboux-Durand
alain.lamboux-durand@univ-fcomte.fr
Université de Franche-Comté

Introduction: automatisation dans l'audiovisuel

Les besoins de rationalisation des coûts et des moyens de production en Europe et dans le monde amènent les services de médias audiovisuels à avoir de plus en plus recours à de l'automatisation au cours de leurs phases de production comme de post-production. La notion d'automatisation fait en elle-même débat dans la mesure où elle est très mal définie : la plupart des dictionnaires indiquent un laconique « Fait d'automatiser ». Il serait plus juste de la qualifier de « stratégie qui vise, par l'emploi de méthodes et de moyens normalisés et optimisés de technologies, à la réduction de l'intervention de l'Homme dans la réalisation d'une opération ou d'une série d'opérations ». Une tâche automatisée relève ainsi de l'ajout de systèmes informatique, robotique,... dans le but de se substituer à l'Homme – réaliser le montage d'un produit audiovisuel par un algorithme, ou remplacer 3 cadreurs par un seul qui pilote 3 caméras à distance est une tâche automatisée. En ce sens, l'automatisation est actuellement portée par les constructeurs de matériel audiovisuel professionnel à travers des systèmes plus ou moins complexes. La conception de ces outils d'automatisation du « workflow » demeur e partielle et aucun système global unifié n'est actuellement proposé. De ce fait, les chaines de télévision associent équipements traditionnels et automates. Le workflow désigne, dans le jargon de l'audiovisuel, l'intégralité de la chaîne de production de l'idée au produit final archivé au sens de (Leleu-Merviel, 1997).

La production d'images et de sons s'est toujours accompagnée d'instruments plus ou moins élaborés (il est difficile de réaliser un film sans caméra). Cependant, le contrôle de cet outil, la prise de décision quant à son utilisation, change d'entité avec l'automatisation. Le matériel est le même, mais son utilisation est gérée, commandée par un automate. Selon (Arendt, 1958), l'action d'un agent humain est imprévisible, car elle est inextricable d'un réseau de relations humaines. Sans pour autant lui enlever son appellation d'artefact, il est nécessaire de s'interroger sur la nature d'un produit qui n'émerge plus uniquement des processus sociotechniques, mais sur des règles d'inférences d'une unité de calcul informatique. Aussi se pose la question de l'impact de cette diminution de l'intervention humaine dans la production audiovisuelle sur sa réception par le spectateur. Un produit audiovisuel est un artefact à but communicationnel, créé par des humains pour des humains. En effet, les choix des acteurs – au sens de (Latour, 2005) – de la création d'une l'œuvre audiovisuelle ont un impact (variable mais bien présent) sur ce dernier. Si le type de produit (fiction, publicité, émission de divertissement, journal télévisé,...) va influencer les modalités de réception de ce produit – ce que (Jauss, 1978) appelle l'Horizon d'attente – les paramètres de conception et de réalisation du produit doivent eux aussi rentrer en compte. Il est probable que l'automatisation des systèmes de production audiovisuelle, entraînant un formatage plus fort du produit à créer, impacte le sens construit par le spectateur. Toutefois demeure la question de l'importance de chaque maillon de la chaîne de production : quelle est la signifiance apportée par l'humain vis-à-vis d'un automate dans la production de l'œuvre ?

Méthodologie: analyse des processus sociotechniques et étude statistiques

L'analyse des processus sociotechniques en régie de production et des acteurs du programme sur le plateau, comparée aux nouvelles méthodes de travail d'un « workflow » automatisé, permet de relever des différences dans la conception du produit. Ces différences, dans l'esprit des travaux de (Latour, 2005) seront soulignées par les agents contribuant à la création du produit eux-mêmes. Ainsi, la méthode permet de déceler des finesses qu'une simple comparaison des tâches à effectuer ne permet pas. Par la suite la mise en parallèle avec un test d'impact de plusieurs versions d'un même événement télévisuel (traditionnelle et automatisée) sur le public permet au chercheur de faire le lien entre modification d'un réseau d'acteurs de la chaîne de produit et sens construit par le spectateur.

Dans le cadre de cette étude, une captation multi-caméra et à plusieurs systèmes de captation a été mise en place. Le sujet filmé relevait de parenthèses musicales lors d'une remise de diplômes à l'Université de Valenciennes. Trois systèmes ont été utilisés en parallèle :

- Un système "témoin", mobilisant une régie professionnelle d'une vingtaine de personnes, étudiants de master audiovisuel de l'Université de Valenciennes ;
- Un système "semi-automatisé", utilisant une régie miniaturisée pour une personne (réalisateur professionnel) ainsi que des caméras robotisées pilotées à distance par un second opérateur ;
- Un système "mono", consistant en une caméra filmant le sujet en plan large fixe.

Par la suite, les encadrants de l'équipe “standard” et les intervenants de l'équipe “semi-automatisée” ont été interrogés sur la préparation et le déroulement de la captation dans leurs régies respectives. Ces résultats ont été croisés avec une étude statistique de l'utilisation des différentes caméras et des réalisations sur les séquences filmées, reconstitués à partir de l’enregistrement des réalisations ainsi que de chaque caméra individuellement.

	Témoin	Semi-automatisé
Nombre de caméras	7	3
Minimum de plans par caméras	1	3
Maximum de plans par caméras	22	6
Ecart type	7.06	1.53
Moyenne	8.71	4.33

Illustration 1: Éléments statistiques issus de l'analyse des réalisations

Les premiers résultats ont souligné une préparation plus forte de la réalisation et un formatage des angles de caméras pour la régie semi-automatisée par rapport à la régie standard. Du fait que les cadres soient mémorisés par le système de contrôle des caméras robotisées, ces derniers sont peu appelés à évoluer au fil du temps, simplifiant les échanges entre réalisateur et opérateur des caméras. À l’inverse, les valeurs de plans du système témoin sont plus fluctuantes. Malgré cette différence, les valeurs employées par les deux systèmes sont globalement identiques quant au découpage de l’espace pour le réalisateur.

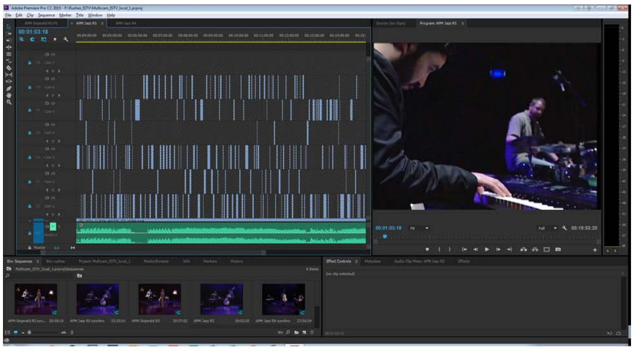


Illustration 2: Capture d'écran de la reconstitution des réalisations sur un logiciel de montage

Suites prévues: tests d'impact et nouvelles captations

Dans la continuité, un test d'impact des séquences filmées sur le public sera effectué selon une adaptation de la méthode Média-Repères (Labour, 2011), qui a pour objectif de diagnostiquer l'impact d'un extrait vidéo sur le public par l'identification des construits de sens chez la personne. La méthode permettra de distinguer des registres de sens construit chez le spectateur afin de discriminer les apports de chaque système de captation en fonction de leur impact sur le public. En particulier, une partie du protocole se fondra sur l'analyse des émotions du spectateur lors du visionnage via le procédé SYM (Yvart et al., 2016). Les résultats issus de cette expérimentation prochaine seront présentés lors du congrès.

Des améliorations sont prévues pour la prochaine étude des systèmes de captations, où la récolte de traces se verra renforcée afin de d'améliorer la qualité de la réminiscence des acteurs ainsi que de d'identifier des nouveaux paramètres quant à l'impact de l'automatisation chez le spectateur. De plus, il sera possible de tester les différents systèmes de captation successivement (ce qui évite les contraintes de placements des multi captations) et pourra se faire sur les mêmes équipements, mais utilisés soit par des humains, soit pilotés par des machines.

Bibliography

Arendt, H. (1958). *The Human Condition*. Chicago: University of Chicago Press.

Jauss, H. R. (1978). *Pour Une Esthétique de La Réception*. TEL. Paris: Gallimard.

Labour, M. (2011). MEDIA-REPERES. Une méthode pour l’explicitation des construits de sens au visionnage Université de Valenciennes et du Hainaut-Cambrésis. Lille Nord de France Habilitation à diriger des recherches.

Latour, B. (2005). *Reassembling the Social, An Introduction to Actor-Network-Theory*. Oxford University Press.

Leleu-Merviel, S. (1997). *La Conception En Communication, Méthodologie Qualité*. Paris: Editions Hermès.

Yvart, W., Delestage, C.-A. and Leleu-Merviel, S. (2016). SYM: Toward a new tool in user’s mood determination. *Proceedings Emovis 2016*. Sonoma, CA. A venir.

Using Big Data Techniques For Searching Digital Archives: use cases in Digital Humanities

Janet Delve
Janet.Delve@port.ac.uk
University of Portsmouth, United Kingdom

Sven Schlarb
sven.schlarb@ait.ac.at
Austrian Institute of Technology, Austria

Rainer Schmidt
rainer.schmidt@ait.ac.at
Austrian Institute of Technology, Austria

Richard Healey
Richard.Healey@port.ac.uk
University of Portsmouth, United Kingdom

Background

The background for this paper is work in progress in E-ARK: an EC FP7 pilot B project[1] having as its main objective the creation of an open source, digital archiving system with attendant standards and tools to be deployed in seven pilot instances. Hence practical application is at the heart of the project, which is led by archivists, researchers, SMEs, digital preservation / archiving membership organizations and government home offices, who together seek to fill the current digital archiving lacuna. E-ARK is a wide-ranging project: we are taking and integrating existing best practices into a digital archiving system, so that it is suitable not only for national archives and government agencies, but also for regional, local, business and research archives of many shapes and sizes. A legal study taking account of varying national legal directives delineates how the archiving system can be deployed against a pan-European backdrop.

At the heart of the system is the OAIS standard (OAIS), with data arriving at an archive via Submission Information Packages (SIPs), which are then stored in the archive as Archival Information Packages (AIPs) and subsequently retrieved upon access as Dissemination Information Packages (DIPs). E-ARK is also addressing an eclectic range of sources: both structured and unstructured data, atomic and complex, including records from Electronic Records Management Systems (ERMSs); databases; geodata; and computer files.

The project stakeholders are similarly also drawn from a wide pool of varied users, and include public administrations, public agencies, public services, citizens, researchers and business. Re-use of information is a key project objective, and we are employing the latest Big Data tools / techniques / architecture such as Hadoop and Lily to present users with innovative access methods, such as the data mining showcase using geodata. We are also building on techniques used in creating an Oracle data warehouse of US 1880 census data (Healey and Delve, 2007).

Although E-ARK is being spearheaded by national archives, it is a key objective to be relevant and useful to a broad church, and to that end the paper should be of interest to many in the Digital Humanities community, not just archives, so we will be including use cases tailored to this end.

Scope of the Paper

E-ARK has conducted a GAP analysis that identifies user requirements for access services, which are described in project report D5.1 (Thirifays et al., 2014). The study investigates the current landscape of archival solutions regarding currently-available access components and identifies gaps and requirements from the perspective of national archives and third party users, as well as content providers. This report has identified a major gap in the identification process, where users browse and search collections to identify material of potential interest. It states that a lack of comprehensive metadata that is available and indexed in finding aids, compromises their performance and efficiency, which directly impacts the user experience and the user's access to the archival holdings in their entirety.

To fill this gap, work on the E-ARK Faceted Query Interface and Application Programming Interface (API) aims to establish a scalable search infrastructure for archived content. It is important to note that here we are not working with the whole content ecosystem of an archive, but instead concentrating only on indexing and searching of the born-digital E-ARK Information Packages (IPs). The goal is not to replace existing systems but to augment these components (like available catalogues) with a "content repository" that can be searched based on a full text index. This content repository concentrates on search and access based on the content (ie. data/files) contained within an AIP rather than selected metadata elements. The reference implementation employs scalable (cluster) technology, as scalability issues must be taken into account when operating a detailed content-based search facility. A major task in the context of the reference implementation is the development of a faceted query interface for searching archived content that can be utilized directly by end-users or integrated with other software components like archival catalogues.

Work on *Query and Indexing* concerns the configuration and generation of a repository index that holds detailed information on the archives' digital holdings. For developing a reference implementation, it is important to provide a solution that is flexible and configurable with respect to a range of requirements. The exact configuration of the faceted search interface will be driven by requirements of the access components (like DIP creation)[2] as well as individual institutional requirements and content specific aspects. As a consequence, the reference implementation developed within E-ARK must provide a configurable query interface that should be accessible via a service API. This API can be used through a web interface and/or an access component for searching the repository based on a full text index.

The reference implementation integrates this query API with a repository implementation, which in turn provides access to an application layer via its access API. The application layer implemented in E-ARK develops end user components for search, access, and display of archived records. Figure 1 provides a conceptual overview of the architecture and workflow supported by the reference implementation.

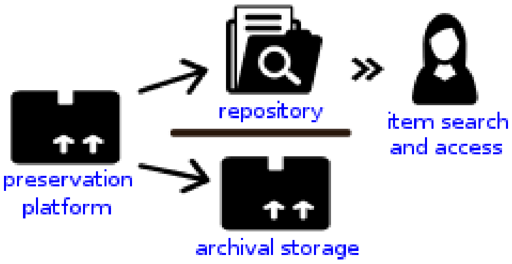


Figure 1:[3] IPs are received and processed by an ingest system like the ESSArch Preservation Platform[4]. As part of the ingest workflow, the information packages are written to an archival storage medium. In addition[5], these packages are also ingested into a content repository that provides basic data management capabilities and search. The created repository records are indexed and can be queried by an end-user application through a service interface. At the repository level, random access is provided on an item/file based level.

The goal of the E-ARK Faceted Query Interface and API is the establishment of a reference service that enables application components to search through the entire archived data and to link the applications with the data management layer[6] (provided through the content repository). The reference implementation will concentrate on data management functionality that supports search, access, and data mining (like providing a CRUD API and support for versioning). The implementation of a fully-fledged archival data management system, however, is out of focus for the reference implementation.

The search functionality is provided by an indexing infrastructure which generates a full-text index for data being ingested into the data management component (ie. the content repository). The goal is to enable end users to efficiently search archival records based on different aspects (or facets) extracted from the archived data and metadata. The search index includes enclosed archival descriptions (metadata) but most crucially the archived data itself (e.g. based on extracted text portions) and generated technical metadata (like file format information).

The employed indexing techniques are not intended to provide a finding aid based on archival metadata, as for example provided by archival cataloguing systems. The intention of the E-ARK Faceted Query Interface and API is to provide a complementary service that takes advantage of information retrieval techniques like full text indexing, faceted search, and ranking to improve the search through archived data. The indexing workflow is however configurable and able to extract specific information from the archival metadata. This flexibility can for example be utilized to develop specific search facets and/or to handle information related to data confidentiality.

The Faceted Query Interface and API are being developed as part of the E-ARK reference implementation which builds upon a scalable technology stack. The intention is to provide an archiving and search solution that works for different payloads. The reference implementation can therefore be scaled from a single host out to a cluster deployment that is capable of maintaining large volumes of data, e.g. in the magnitude of hundreds of terabytes of archived data organized in hundreds of millions of repository records. The indexing infrastructure is however intended to be deployed next to established archiving systems in order to extend the functionality of the available finding aids. The intention is not to replace the existing systems but rather to extend these infrastructures.

The final paper proposes to add further details of deploying the above scenario with use cases making use of geographic data integrated with the peripleo tool from the Pelagios project[7]. We will describe the implementation of a complete archival workflow that includes conversion procedures necessary to support text-based search as well as geographic information retrieval and spatial browsing. We will also show how Big Data techniques such as denormalisation and dimensional modelling used in creating the AIPs can facilitate the discovery methods we outline.

Bibliography

Healey, R. and Delve, J. (2007). Integrating GIS and Data Warehousing in a Web Environment: A Case Study of the US 1880 census, *International Journal of Geographical Information Science*, **21**(6): 603-24.

Thirifays, et al. (2014). GAP report between requirements for access and current access solutions. http://www.eark-project.com/resources/project-deliverables/3-d51-e-ark-gap-report (accessed 29 October 2015).

Thirifays, et al.. (2015). E-ARK DIP draft specification. http://www.eark-project.com/resources/project-deliverables/31-d52 (accessed 5 March 2016).

OAIS. (2012). Reference Model for an Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0m2.pdf (accessed 29 October 2015).

Notes

[1] E-ARK is funded by the European Commission's FP7 PSP CIP Pilot B Programme under Grant Agreement no. 620998.

[2] The specific access component requirements are being currently defined in E-ARK and are already partially described in (Thirifays et al., 2015), the report D5.2 "E-ARK DIP draft specification"

[3] Icons made by Freepik

[4] http://www.essarch.org/

[5] The full-text search and access component developed in E-ARK does not replace an existing archival system (like catalogues) but can be utilized to augment these systems.

[6] Here, data management refers to functionality provided by the content repository introduced by the E-ARK infrastructure, which is intended to augment the existing archival ecosystem.

[7] https://github.com/pelagios/peripleo

SAMEBibl: Sistema Automático de Migración a Europeana para Bibliotecas.

M Luisa Diez-Platas
ml.diezplatas@gmail.com
Universidad Pontificia de Salamanca en Madrid, Spain

Paloma Centenera-Centenera
paloma.centenera@gmail.com
Universidad Pontificia de Salamanca en Madrid, Spain

Introducción

Europeana, como biblioteca digital europea, se considera como el medio de referencia para compartir el extenso patrimonio europeo de distintas categorías y del que son depositarias instituciones de diversa índole, ya sean públicas o privadas. Desde la puesta en marcha de esta biblioteca han estado presentes instituciones relevante y entre ellas bibliotecas como la *British Library*.

Para las bibliotecas también es importante su integración en Europeana ya que las dota de una visibilidad imposible de imaginar en los ámbitos más restringidos en los que se encuentran localizadas las bibliotecas y les permite compartir su extenso patrimonio.

El proceso de adaptación de la información, tradicionalmente estructurada en bases de datos relacionales, al modelo de datos abiertos de Europeana (*Europeana Data Model*, EDM) es excesivamente complejo, difícil de automatizar y requiere de la intervención de expertos. La complejidad estriba fundamentalmente en la brecha semántica que existe entre la forma en la que las bibliotecas almacenan y estructuran su información, en general, y la forma de representación que requiere *Europeana*. Este modelo de datos es una especificación conceptual basada en etiquetas semánticas que se integran en los datos de forma que datos y metadatos forman un todo.

Las mayoría de las bibliotecas cuentan en la actualidad con una base de datos relacional, un modelo de datos de representación en el que ya se muestran detalles que trascienden los conceptos del mundo modelado, especificando características más relacionadas con la implementación de conceptos y relaciones.

Es necesario, pues, extraer la semántica conceptual del modelo relacional y trasladarla al modelo que Europeana define y exige para poder integrar en ella la biblioteca y dar visibilidad a su patrimonio.

SAMEBibl, automatización del proceso de migración a Europeana

SAMEBibl , nace con el objetivo de cubrir el proceso completo de migración de una biblioteca para su integración en la Europeana. Este proceso, que se muestra en Figura 1, parte del modelo origen de representación de la información y lo transforma en el modelo destino.

La semántica de los dos modelos, origen y destino, es cualitativamente distinta y por tanto, el proceso requiere que se conciban modelos conceptuales intermedios que evolucionan progresivamente hasta obtener el modelo necesario para la integración de los datos en Europeana.

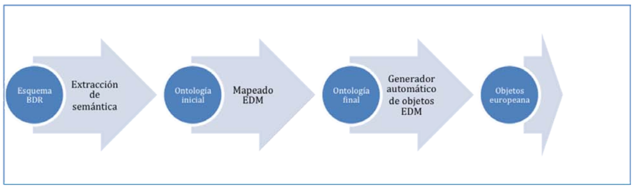
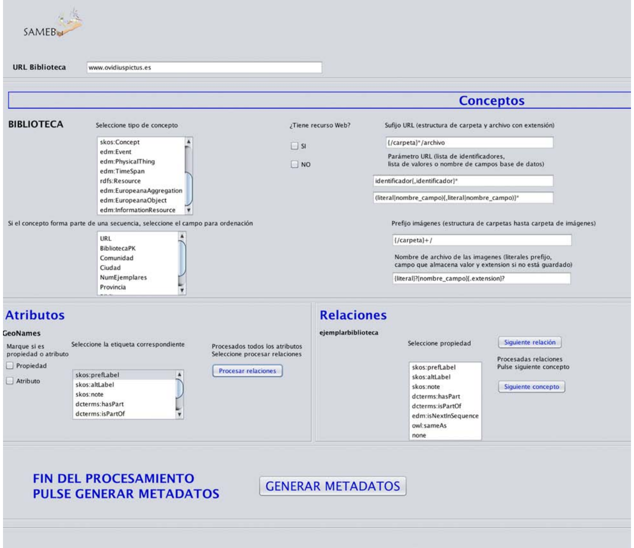


Figura 1



URL Biblioteca: www.ovidiuspictus.es

Conceptos

BIBLIOTECA

Seleccione tipo de concepto

¿Tiene recurso Web?

□ SI
□ NO

Sufijo URL (estructura de carpeta y archivo con extensión): {/carpeta}*/archivo

Parámetro URL (lista de identificadores, lista de valores o nombre de campos base de datos): identificador{,identificador}*

(literal|nombre_campo){,literal|nombre_campo)}*

Si el concepto forma parte de una secuencia, seleccione el campo para ordenación

Prefijo imágenes (estructura de carpetas hasta carpeta de imágenes): {/carpeta}+/

Nombre de archivo de las imagenes (literales prefijo, campo que almacena valor y extension si no está guardado): {literal}?{nombre_campo}{.extension}?

Atributos

GeoNames

Marque si es propiedad o atributo

□ Propiedad
□ Atributo

Seleccione la etiqueta correspondiente

Procesados todos los atributos
Seleccione procesar relaciones

Relaciones

ejemplarbiblioteca

Seleccione propiedad

Procesadas relaciones
Pulse siguiente concepto

FIN DEL PROCESAMIENTO
PULSE GENERAR METADATOS

Figura 2. Pantalla interfaz Samebibl

SAMEBibl automatiza casi todo el proceso de migración, aunque se requiere la intervención del experto de la biblioteca con el objeto de establecer las correspondencias mas convenientes entre los componentes del modelo original y el modelo de *Europeana*, permitiendo asimismo la elección de la información que se considera más oportuno migrar. La interacción del experto con el sistema se realizará de forma amigable, presentando la información que se quiere migrar de forma independiente a su representación interna en el modelo que se ha definido, es decir, no es necesario que cuente con conocimientos técnicos en lo que a repositorios y datos abiertos se refiere.

La aplicación desarrollada consta de tres componentes principales:

- Un traductor dirigido por la sintaxis que extrae la semántica del modelo relacional a partir de las estructuras sintácticas que se identifican en el esquema de la base de datos, y que dan lugar a un modelo conceptual intermedio.
- Un módulo que permite establecer las correspondencias, es decir, definir el "mapeado", entre cada uno de los conceptos y relaciones del modelo conceptual extraído por el traductor y el modelo de Europeana. Para ello se requiere la intervención del experto de la biblioteca.
- Un traductor que, a partir de la información obtenida de los registros de la base de datos, la información facilitada en el mapeado por el experto y los elementos del modelo conceptual, genera los metadatos basados en *EDM* en un formato XML/RDF necesario para poblar el repositorio que será recolectado por *Europeana*.

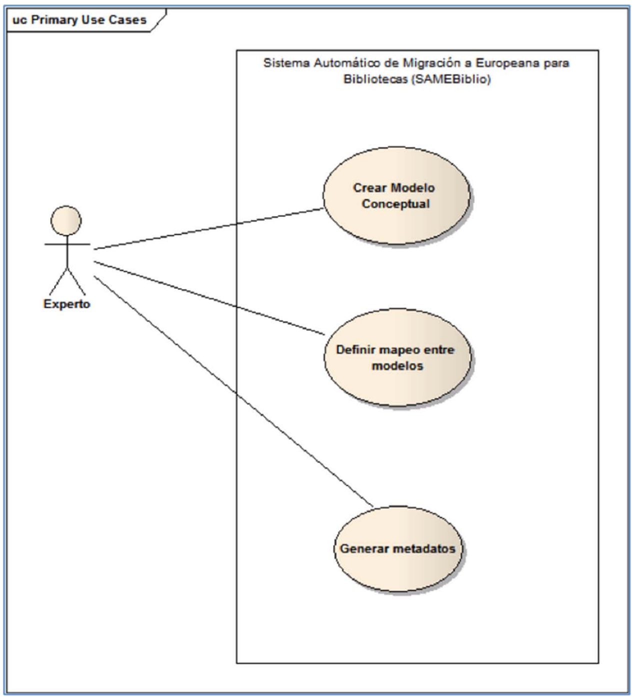

uc Primary Use Cases
Sistema Automático de Migración a Europeana para Bibliotecas (SAMEBiblio)
Experto
Crear Modelo Conceptual
Definir mapeo entre modelos
Generar metadatos

Figura 3. Casos de uso Samebibl

El proceso de migración no sólo presenta complejidad interna. La cantidad de información manejada y transformada puede ser de tamaño considerable y por tanto, requerir de tiempo para su realización. Además, la información puede ser cambiante o se puede añadir información nueva. Es importante, por tanto que SAMEBibl dote de persistencia a la información que se ha generado de la persistencia necesaria para permitir su modificación y actualización. Debido a esto, la información relativa a los modelos y sus correspondencias, es almacenada para permitir tanto su actualización y mejora, con la revisión o incorporación de nuevas reglas de mapeado tras la incorporación de nuevos elementos semánticos.

Esta característica facilita el mantenimiento automático del repositorio que incorporará los cambios surgidos en el entorno operacional de la base de datos relacional.

SAMEBibl , por tanto, es una herramienta que cubre una necesidad detectada en muchas bibliotecas,. Esta necesidad es mayor en aquellas bibliotecas que no tienen recursos suficientes como para contar con expertos técnicos que realicen el proceso de transformación del modelo de su base de datos relacional a *Europeana*., y esta es la situación en la que se encuentran la mayoría de las bibliotecas europeas que son de menor envergadura pero que son depositarias de un gran patrimonio. Sus tres características más relevantes son:

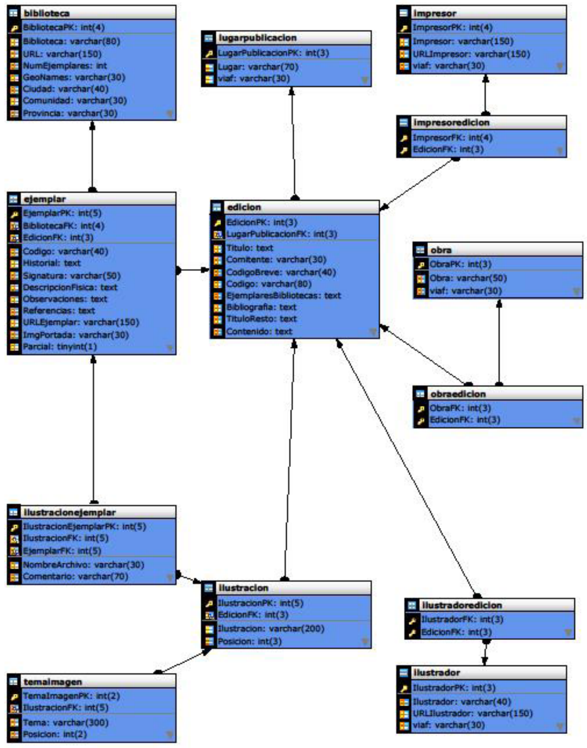
- la automatización completa de las funciones de extracción de la semántica conceptual y generación de metadatos. Estas dos funciones abarcan los elementos técnicamente más complejos del proceso de migración. Su total automatización aísla al usuario de esta complejidad.
- la simplicidad y naturalidad en la operación de la funcionalidad de mapeado, al contar con una interfaz sencilla que permite al usuario de la biblioteca establecer las correspondencias semánticas entre los modelos de acuerdo a sus criterios de forma intuitiva y amigable.
- la flexibilidad del modelo adoptado en la representación del conocimiento relativo a los modelos y sus correspondencias. Esto permite proporcionar ante una herramienta de propósito general que puede ser empleada por cualquier biblioteca basada en un modelo relacional. Asimismo hará muy sencilla la adaptación de la herramienta en el futuro para soportar la migración a nuevos modelos de metadatos semánticos distintos de *EDM*.

Aplicación de *SAMEBibl* a la Biblioteca Digotal Ovidiana

SAMEBIbl ya ha sido aplicada en la *Biblioteca Digital Ovidiana (desde ahora BDO)* (*www.oviduspictus.es*) dedicada a la obra ilustrada de Ovidio y que alberga los datos y las imágenes de todos los ejemplares de las ediciones impresas entre los siglos XV al XIX que se encuentran en las bibliotecas españolas. Es un proyecto que se esta llevando a cabo en varias fases.

Con el objeto de poder compartir la información de la BDO en Europeana, se ha utilizado SAMEBibl para llevar a cabo la migración .

El modelo lógico relacional de la base de datos de BDO que se muestra en Figura 3 es un modelo complejo ya que refleja la rica semántica y estructuración de la información almacenada en la misma



Tras el proceso de migración se obtienen los objetos semánticos correspondientes a los diferentes conceptos identificados a partir de la semántica extraída del modelo lógico relacional como se muestra en la Figura 5.

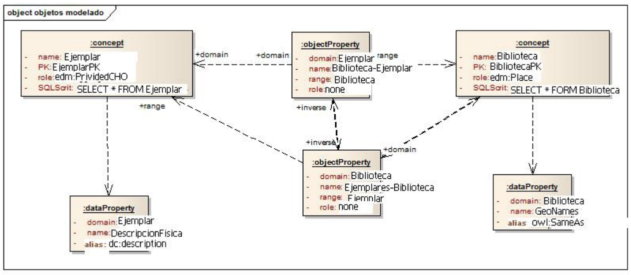


Figura 5. Objetos generados para los conceptos Ejemplar y Biblioteca

A partir del modelo de objetos generado se generan los metadatos y el archivo RDF/XML que permite la integración de la información en el repositorio., tal y como se muestra en Código 1

```
<?xml version=”1.0” encoding=”UTF-8”?>
<rdf:RDF xsi:schemaLocation = “http://www.
    w3.org/1999/02/22-rdf-syntax-ns#http://www.europeana.
    eu/schemas/edm/EDM.xsd” >
<edm:Place rdf:about = “http://www.ovidiuspictus.es/
    biblioteca/27” rdf:resource = “http://bibliotecas.usal.
    es/?q=biblioteca/general-historica” >
<skos:prefLabel>
Biblioteca General Histórica de la Universidad.
</skos:prefLabel>
<dcterms:isPartOf>
Ciudad: Salamanca
</dcterms:isPartOf>
<dcterms:isPartOf>
Provincia: Salamanca
</dcterms:isPartOf>
<dcterms:isPartOf>
Comunidad autónoma: Castilla León
</dcterms:isPartOf>
</edm:Place>
<edm:ProvidedCHO rdf:about = “http://www.ovidiuspictus.
    es/edicion/137” >
<dc:title>
PUBLII <br />OVIDII NASONIS <br />OPERA OMNIA <br
    />IV. VOLUMINIBUS COMPREHENSA; <br />CUM
    INTEGRIS <br />JACOBI MICYLLI, HERCULIS CIOFA-
    NI, <br />ET DANIELIS HEINSII <br />NOTIS, <br
    />ET <br />NICOLAI HEINSII <br />CURIS SECUN-
    DIS, <br />ET <br />ALIORUM IN SINGULAS PARTES,
    <br />PARTIM INTEGRIS, PARTIM EXCERPTIS, <br
    />ADNOTATIONIBUS, <br />CURA ET STUDIO <br
    />PETRI BURMANNI, <br />QUI ET SUAS <br />IN
    OMNE OPUS NOTAS ADIECIT. <br /><br />(marca
    de impresor) <br /><br /><i>AMSTELODAMI, <br
    /></i>APUD FRANCISCUM CHANGUION. <br
    />M.D.C.C.XXVII. <br />
</dc:title>
<dc:identifier>OO.Changuion.Ams.1727.t1 </dc:identifier>
<dc:identifier>Obras Completas.Burmannus.Changuion.
    Ámsterdam.1727.t1 </dc:identifier>
<dc:tableOfContents>
Anteportada; <br/>frontispicio; <br />portada de las obras
    completas; ENTRE PÁGINAS 322 Y 323 (COMENTA-
    RIO AL EJEMPLAR) <br />f. []4 dedicatoria al príncipe
    Eugenio de Saboya; <br />f. []4 v. poema dedicatorio; <br
    />f. *** Prefacio fechado en 1726; <br />p. 1 <i>Heroidas
    </i>; <br />p. 321 anteportada <i>Amores </i>; <br
    />p. 323 epigrama de Ovidio sobre sus <i>Amores </i>;
    <br />p. 325 comienza los <i>Amores </i>; <br />p. 537
    anteportada <i>Arte de amar </i>; <br />p. 539 comien-
    za el <i>Arte de amar </i>; <br />p. 713 anteportada
    <i>Remedios contra el amor </i>; <br />p. 715 comienza
    el libro de los <i>Remedios contra el amor </i>; <br />p.
    766 <i>Sobre la cosmética del rostro femenino </i>; <br
    />p. 777 <i> Haliéutica </i>; <br />p. 793 anteportada de
    <i>Consolatio ad Liviam </i>, <i>Nux y Heroidas </i>de
    Angelo Sabino; <br />p. 795 comienza la <i> Consolatio
    ad Liviam </i>; <br />p. 829 comienza la <i> Nux </i>;
    <br />p. 844 comienzan las <i>Heroidas </i>de Angelo
    Sabino;
</dc:tableOfContents>
<dc:description>Comentario: Notas de Jacobus Micyllus,
    Hercules Ciofano y Nicolas Hensius. </dc:description>
<dc:description>Estructura disposición: Texto latino de an-
    cho de página con versos numerados y notas filológicas al
    pie, organizadas en dos columnas. </dc:description>
<dc:description>Ilustraciones: La edición lleva un frontispicio
    calcográfico diferente en cada uno de los tres primeros
```

```
tomos. En este primero se representa a Ovidio, coronado de laurel y con el pecho atravesado por una flecha que le ha lanzado Cupido, que vuela por encima de él. El poeta, que está sentado, escribiendo en una tablilla, está siendo coronado por una musa que además sostiene una rama y tiene una lira y una flauta a sus pies. Frente al poeta se encuentra la Fama, representada como una dama con túnica y manto, tocada con una diadema, que lleva una tuba en la mano derecha mientras con la izquierda con una especie de puntero le indica a Ovidio lo que debe escribir. La escena tiene lugar al borde de una corriente de agua surcada por una pareja de cisnes. En el árbol que hay detrás del grupo cuelga una cartela que reza <i>P. OVIDII NASONIS OPERA OMNIA. </i><br />El grabado va firmado por el diseñador y el grabador: Bernard Picart ( <i>B Picart inv. </i>) y Matthys Pool ( <i>M. Pool Sculp) </i>. <br />Esta misma edición vio la luz en Ámsterdam en el mismo año en casa de tres impresores distintos: Changuion, Westein & Smith y Janssonius Waesbergius. (Cf. ediciones en la <i>Biblioteca Digital Ovidiana </i>). <br />Un ejemplar de esta edición se encuentra completamente digitalizado en la Biblioteca Nacional http://bdh.bne.es/bnesearch/detalle/bdh0000052654
</dc:description>
<dcterms:created>1727 </dcterms:created>
<edm:realize rdf:resource = "http://www.ovidiuspictus.es/ejemplar/214" />
<dc:contributor rdf:resource = "http://www.ovidiuspictus.es/ilustrador/22" />
<dc:contributor rdf:resource = "http://www.ovidiuspictus.es/ilustrador/40" />
<dc:subject rdf:resource = "http://viaf.org/viaf/312263158/" >Arte de amar </dc:subject>
<dc:subject rdf:resource = "http://viaf.org/viaf/303708953/" >Remedios contra el amor </dc:subject>
<dc:subject rdf:resource = "http://viaf.org/viaf/184346141/>Amores </dc:subject>
<dc:subject>Sobre la cosmética del rostro femenino </dc:subject>
<dc:subject>Haliéutica </dc:subject>
<dc:subject>Consolatio ad Liviam </dc:subject>
<dc:subject>Nux </dc:subject>
<dc:contributor rdf:resource = "http://www.ovidiuspictus.es/impresor/65" />
<dcterms:spatial rdf:resource = "http://www.geonames.org/2759794/amsterdam.html" >Ámsterdam (Amstelodami) </dcterms:spatial>
</edm:ProvidedCHO>
<edm:WebResource dc:about = "http://www.ovidiuspictus.es/visualizacionejemplar.php?clave=214%20&%20clave1=OO.BGH.Ams.1727a.t1" >
<dc:format>IMAGE </dc:format>
<dcterms:created>Biblioteca Digital Ovidiana </dcterms:created>
<dcterms:rights>Biblioteca Digital Ovidiana </dcterms:rights>
<dcterms:hasPart rdef:resource = "htt://wwww.ovidiuspictus/images/images/OO.BGH.Ams.1727a.t1/OO.BGH.Ams.1727a.t1.1" />
</edm:WebResource>
<edm:ProvidedCHO dc:about = "http://www.ovidiuspictus.es/ejemplar/214" >
<dc:title>OO.BGH.Ams.1727a.t1 </dc:title>
<dc:type>BOOK </dc:type>
<dc:identifier>BG/ 35111 </dc:identifier>
<dc:provenance rdf:resource = "http://www.ovidiuspictus.es/biblioteca/27" />
</edm:ProvidedCHO>
<ore:Aggregation dc:about = "http://www.ovidiuspictus.es/ejemplarAg/214" >
<edm:isShownAt rdf:resource = "http://www.ovidiuspictus.es/visualizacionejemplar.php?clave=214%20&%20clave1=OO.BGH.Ams.1727a.t1" />
<edm:dataProvider>Biblioteca Digital Ovidiana </edm:dataProvider>
<edm:AggregatedCHO rdf:resource = "http://www.ovidiuspictus.es/ejemplar/214" />
<edm:hasView rdf:resource = "http://www.ovidiuspictus.es/ejemplar/214" />
</ore:Aggregation>
<edm:WebResource dc:about = "http://wwww.ovidiuspictus/images/images/OO.BGH.Ams.1727a.t1/OO.BGH.Ams.1727a.t1.1" >
<dc:format>IMAGE </dc:format>
<dcterms:created>Biblioteca Digital Ovidiana </dcterms:created>
<dcterms:rights>Biblioteca Digital Ovidiana </dcterms:rights>
<edm:aggregatedCHO rdef:resource = "http://www.ovidiuspictus.es/visualizacionejemplar.php?clave=214%20&%20clave1=OO.BGH.Ams.1727a.t1" />
</edm:WebResource>
<edm:ProvidedCHO dc:about = "http://www.ovidiuspictus.es/ilustracionejemplar/1696" >
<edm:isShownAt rdf:resource = "http://wwww.ovidiuspictus/images/images/OO.BGH.Ams.1727a.t1/OO.BGH.Ams.1727a.t1.1" />
<dc:title>Frontispicio </dc:title>
<dc:description>Ovidio, herido de amor por Cupido, escribiendo poesía, asistido por una Musa que lo corona y la Fama que le indica qué escribir </dc:description>
</edm:ProvidedCHO>
<edm:Agent dc:about = "http://www.ovidiuspictus.es/ilustrador/22>" >
<rdaGr2:professionOrOccupation >Ilustrador </rdaGr2:professionOrOccupation>
<skos:prefLabel>Pool, Mattys (1670-1732) </skos:prefLabel>
<edm:hasMet rdf:resource = "http://viaf.org/viaf/13247568/" />
<owl:sameAs rdf:resource = "http://thesaurus.cerl.org/record/cnp01001580" ></owl:sameAs>
</edm:Agent>
<edm:Agent dc:about = "http://www.ovidiuspictus.es/ilustrador/40>" >
<rdaGr2:professionOrOccupation >Ilustrador </rdaGr2:professionOrOccupation>
<skos:prefLabel>Bernard Picart </skos:prefLabel>
<edm:hasMet rdf:resource = "http://viaf.org/viaf/64010408/" />
```

```
<owl:sameAs rdf:resource = “http://thesaurus.cerl.org/record/
    cnp01318163” />
</edm:Agent>
<edm:Agent dc:about = “http://www.ovidiuspictus.es/mpre-
    sor/65>” >
<rdaGr2:professionOrOccupation >Impresor </
    rdaGr2:professionOrOccupation>
<skos:prefLabel>Changuion, François </skos:prefLabel>
<edm:hasMet rdf:resource = “http://viaf.org/viaf/19854560/”/>
<owl:sameAs rdf:resource = “http://thesaurus.cerl.org/record/
    cni00012233” />
</edm:Agent>
</rdf:RDF>
```

Conclusiones

- El proceso de migración de una biblioteca digital basada en un modelo de datos relacional a Europeana entraña una gran dificultad. Esto es debido en gran parte a la gran brecha existente entre ambos modelos, el primero de naturaleza más lógica y orientada a la representación de los datos y el segundo conceptual. Esta se constata, entre otras cosas, por el reducido número de bibliotecas que se han integrado en Europeana.
- Faltan herramientas para las bibliotecas puesto que las existentes no cubren todo el proceso, son difíciles de emplear y, o bien no son específicas para Europeana , o son muy costosas.
- El software SAMEBibl desarrollado, integra y automatiza, en la medida de lo posible, todos los procesos de la metodología. Esta herramienta lleva a cabo un proceso de traducción sintáctica y semántica del modelo origen con el establecimiento de las relaciones y reglas de transformación que da lugar a un archivo en un formato interpretable por un repositorio.
- La metodología y la herramienta cumplen con los objetivos planteados como demuestra su validación mediante un caso real, la Biblioteca Digital Ovidiana (http://www.ovidiuspictus.es).

Bibliography

Agenjo Bullón, X., and Hernández Carrascal, F. (2011). *Perspectivas europeas en el desarrollo funcional de los sistemas de información: la agregación de datos del europeana data model. FESABID’11. Actas de Las XII Jornadas Españolas de Documentación.*

Aloia, N., Concordia, C., and Meghini, C. (2013). *The Europeana Linked Open Data Pilot Server.* In M. Agosti, F. Esposito, S. Ferilli, and N. Ferro (Eds.), Digital Libraries and Archives, pp. 241–48. Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-35834-0_24

Angjeli, A., Bayerische, M., Chambers, S., Charles, V., Clayphan, R., Deliot, C., et al. (2012). *D5. 1 Report on the alignment of library metadata with the European Data Model (EDM) Version 2.0.* Report, Europeana Project.

An, Y., Borgida, A., and Mylopoulos, J. (2005). *Refining semantic mappings from relational tables to ontologies.* In Semantic Web and Databases pp. 84–90. Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-31839-2_7

Barrasa Rodríguez, J. (2007). *Modelo para la definición automática de correspondencias semánticas entre ontologías y modelos relacionales (phd).* Facultad de Informática (UPM). Retrieved from http://oa.upm.es/4147/

Berners-Lee, T. (2013). *Relational databases on the semantic web.* Retrieved from http://www.citeulike.org/group/17638/article/11988241

Charles, V., Isaac, A., Tzouvaras, V., and Hennicke, S. (2013). *Mapping Cross-Domain Metadata to the Europeana Data Model (EDM).* In T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. J. Farrugia (Eds.), Research and Advanced Technology for Digital Libraries, pp. 484–85. Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-40501-3_68

Concordia, C., Gradmann, S., and Siebinga, S. (2010). Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. *IFLA Journal*, **36**(1): 61–9, Retrieved from http://ifl.sagepub.com/content/36/1/61.short

CVC (2014). *El español en el mundo. Anuario del Instituto Cervantes 2010-2011. Las bibliotecas digitales del siglo siglo XXI. Rafael C. Carrasco Jiménez.* Retrieved September 18, 2015, from

Europeana. (2014). *Europeana Data Model Mapping Guidelines.* Retrieved from http://pro.europeana.eu:9580/documents/900548/60777b88-35ed-4bae-8248-19c3696b81fb

European Commision. (2005). *i2010: Digital libraries.* Retrieved from http://europa.eu/legislation_summaries/information_society/strategies/l24226i_en.htm

Hernando-De-Larramendi, L., Domínguez-Muriel, J., Viedma-Peláez, A., Hernández-Carrascal, F. and Agenjo, X. (2009). *Datos y metadatos: la normalización dinámica de los elementos y de los procesos constituyentes de una Biblioteca Virtual.* Retrieved from http://eprints.rclis.org/14342

López, F.-A. (2013). *Visibilidad e impacto de los repositorios digitales en acceso abierto. De Bibliotecas Y Bibliotecarios. Boletín Electrónico ABGRA, (5).* Retrieved from http://eprints.rclis.org/18940/

Pan, Z. and Heflin, J. (2004). *Dldb: Extending relational databases to support semantic web queries. DTIC Document.* http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA451847

Remedios Melero, E. A. (2009). *The situation of open access institutional repositories in Spain: 2009.* Retrieved September 12, 2015, from http://www.informationr.net/ir/14-4/paper415.html

Ríos-Hilario, A., Martín-Campo, D. and Ferreras-Fernández, T. (2012). Linked data y linked open data: su implantación en una biblioteca digital. El caso de Europeana. *El Profesional de La Información*, **21**(3): 292–97. Retrieved from http://elprofesionaldelainformacion.metapress.com/index/712822300Q7033W3.pdf

Saorín, T., Peset, F. and Ferrer-Sapena, A. (2013). Retrieved July 1, 2014, from http://eprints.rclis.org/21005/

Sequeda, J. F., Tirmizi, S. H., Corcho, O., and Miranker, D. P. (2011). Survey of directly mapping sql databases to the semantic web. *The Knowledge Engineering Review*, **26**(4):

445–86. Retrieved from http://journals.cambridge.org/abstract_S0269888911000208

Vassallo, V. and Piccininno, M. (2012). Aggregating Content for Europeana: A Workflow to Support Content Providers. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides (Eds.), *Theory and Practice of Digital Libraries*, pp. 445–54. Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-33290-6_50

Vega-Ramırez, A., Grangel-González, I., Sáez-Mosquera, I. and Garcıa-Castro, R. (2014). *Procedimiento para la obtención de un modelo ontológico para representar la información contenida en bases de datos.* Retrieved from http://ceur-ws.org/Vol-1219/paper5.pdf

Le Futur Du Livre Électronique En Accès Libre: L'exemple De La Collection "Parcours Numériques"

Michael Eberle-Sinatra
michael.eberle.sinatra@umontreal.ca
Université de Montréal, Canada

Marcello Vitali-Rosati
marcello.vitali.rosati@umontreal.ca
Université de Montréal, Canada

Hélène Beauchef
lnbeauchef@gmail.com
"Parcours numérique", Canada

Cinq constats, déjà soulignés par plusieurs chercheurs (dont Steven Harnad et Jean Claude Guédon, par exemple) nous semblent démontrer l'urgence de développer des politiques institutionnelles pour promouvoir l'accès libre:

1. Cinq multinationales (Wiley, Elsevier, Springer, Taylor and Francis, Sage) gèrent la quasi-totalité de la presse scientifique. Elles prennent depuis des années en otage les contenus produits par les chercheurs;
2. Les contenus publiés sont payés par l'Université (qui paye la recherche des chercheurs);
3. Les bibliothèques sont obligées de payer une deuxième fois ces mêmes contenus;
4. Cela réduit fortement la circulation des contenus scientifiques au seul profit d'entreprises qui ne jouent aucun rôle dans leur production et qui ne contribuent nullement à leur visibilité;
5. Ce système est maintenu en place par le dispositif des évaluations académiques – pour des raisons de CV, le chercheur publie sur des revues "accréditées" quand il pourrait, sans frais ou presque, publier en accès libre.

Pour sortir de ce mécanisme pervers il faut d'abord essayer de faire résistance au modèle imposé par ces multinationales, d'où l'importance des efforts de remettre en question ces politiques d'abonnement, comme l'a fait la Bibliothèque de l'Université de Montréal qui s'est désabonnée de l'offre panier de Wiley en janvier 2014. Mais il faut aussi proposer de nouveaux modèles.

Depuis longtemps, plusieurs initiatives essaient d'aller dans ce sens en faisant la promotion de l'accès libre, entre autres bien évidemment, pour le domaine francophone, érudit.org et revues.org (OpenEdition) tout comme des revues qui remettent en question les problèmes de modèle économique au nom de cette cause. Avec les Presses de l'Université de Montréal nous avons créé une collection en accès libre, "Parcours numériques", au printemps 2014, avec déjà cinq ouvrages parus.

La collection est basée sur l'idée qu'il doit y avoir une complémentarité entre l'édition papier et l'édition numérique, car ces deux formes de publication présupposent des idées différentes de lecture et deux approches différentes à la réception des contenus.

Le livre papier – et on entend ici également le numérique homothétique (epub ou pdf) qui reproduit à l'identique le livre papier sur un support numérique – permet une lecture linéaire. Une thèse peut y être présentée et argumentée de façon complexe. Le lecteur sera capable de suivre de manière linéaire le développement de l'argumentation, de faire un cheminement avec l'auteur en se laissant accompagner d'un bout à l'autre du discours. C'est pour cela que nous avons fait le choix de publier des livres assez courts (120/200 pages) : c'est la longueur adéquate pour présenter une thèse et la démontrer à l'aide d'une argumentation unique et cohérente. Sur un livre papier, on peut passer plusieurs heures pour suivre l'auteur dans tout son raisonnement. Mais il faut que le discours soit linéaire : tout ce qui entraîne une sortie par rapport au chemin principal doit être évacué. L'appareil critique, les références, les parenthèses, les exemples, les images, les statistiques, les détails… tous ces éléments viennent casser la linéarité de la lecture.

L'édition numérique augmentée, en revanche, présuppose une lecture non linéaire, qui procède par approfondissement. On commence par lire un premier texte sur un sujet, on souhaite en approfondir un aspect, puis on glisse sur un contenu qui se trouve ailleurs et qui nous permet d'en savoir plus sur ce qui au départ ne semblait qu'un détail. On navigue, on flâne, le parcours emprunté n'existe pas avant la navigation, il n'a pas été prévu par un auteur ou un éditeur. Ces derniers ont suggéré des pistes, ouvert des portes… puis la navigation est laissée aux lecteurs, à leurs envies, à leur créativité. Dans ce sens, il n'est pas vrai, comme le voudrait Nicholas Carr, que le numérique détruit notre capacité d'attention : il s'agit d'une attention différente, disséminée, qui permet l'approfondissement mais empêche de suivre un discours plus long et unitaire.

La collection "Parcours numériques" offre donc ces deux possibilités de lecture. L'édition en ligne augmentée offre le texte en intégralité, en libre accès, ainsi que toute une série de contenus additionnels qui seront autant de portes prêtes à être ouvertes pour nous emmener ailleurs, vers des approfondissements, des sujets connexes, d'autres formes de contenus, d'autres plateformes, d'autres parcours, qui n'ont pas été nécessairement prévus par l'auteur ou par l'éditeur du livre. C'est également pour cette raison que l'édition en ligne augmentée est gratuite : elle permet d'avoir accès à un univers connecté au livre, un univers qui n'a été créé ni par l'auteur, ni par l'éditeur. Si par la suite le lecteur souhaite se plonger pleinement et complètement dans la thèse de l'auteur afin d'en connaître et d'en comprendre les moindres aspects, il aura probablement davantage envie de le lire de façon linéaire et donc de l'acheter en papier, en epub ou en pdf.

De cette manière, on permet une circulation libre des contenus, un cercle vertueux où est mis en avant le travail de l'auteur grâce aux liens qui sont créés vers d'autres contenus, produits par d'autres. On met simultanément en place un réseau de connaissances et un dialogue. Mais on permet aussi, grâce au papier et à son double numérique homothétique, au discours linéaire de l'auteur d'exister, clair, identifié, reconnaissable. Deux lectures qui ne sont pas en compétition donc, mais qui se complètent.

Notre présentation révèlera les données sur les ventes au cours des deux premières années, les résultats du modèle économique choisi, ainsi que des données sur la plateforme en ligne.

Bibliography

Cohen, D. (2010). Open Access Publishing and Scholarly Values. *Dan Cohen's Digital Humanities Blog*.http://www.dancohen.org/2010/05/27/open-access-publishing-and-scholarly values/.

Guédon, J. C. (2014). Le Libre Accès et la Grande Conversation scientifique. In Sinatra, M., E. and Vitali-Rosati, M. (Eds.), *Pratiques de l'édition numérique*. Montreal: Presses de l'Université de Montréal.

Hall, G. (2008). *Digitize This Book! The Politics of New Media, or Why We Need Open Access Now*. >Minneapolis and London: University of Minnesota Press.

Harnad, S. (2007). The Green Road to Open Access: A Leveraged Transition. In Gacs, A. (Ed.), *The Culture of Periodicals from the Perspective of the Electronic Age*. Paris: L'Harmattanpp, pp. 99–105.

Sinatra, M. E. and Vitali-Rosatti, M. (Eds.) (2014). *Pratique de l'édition numérique*. Montreal: Presses de l'Université de Montréal.

Sinatra, M. E. (2015). Promoting Open Access and Innovation: From Synergies to Le Centre de Recherche Interuniversitaire sur les Humanités Numériques. *Scholarly and Research Communication*, **6**(4). http://www.src-online.ca

Willinsky, J. (2009). *The Access Principle: The Case for Open Access to Research and Scholarship*. The MIT Press.

How IBM Watson Can Help Us Understand Character in Shakespeare: A Cognitive Computing Approach to the Plays

Mattia Egloff
mattia.egloff@unil.ch
University of Lausanne, Switzerland

Davide Picca
davide.picca@unil.ch
University of Lausanne, Switzerland

Kevin Curran
kevin.curran@unil.ch
University of Lausanne, Switzerland

Introduction

The study of an individuals' personality traits is a new line of research that emerged only recently, primarily through the investigation of dialogue systems and weblogs (Gill et al., 2012; Konstantopoulos, 2010; Mairesse and Walker, 2006; Mairesse and Walker, 2007). This paper proposes a novel application for personality classification by leveraging on cognitive computing research and by exploiting the poetic production of theatrical plays. More specifically, this research is circumscribed to the analysis of Shakespeare's tragedies, which offer a rich spectrum of characters for a detailed and in-depth study. This research does not aim at introducing the innovative technological aspects of personality classification in cognitive computing but rather at employing such technology in the study of English literature, and analyse some implications that arise from the results.

Before presenting the core of this research, we outline the psychological theory of the "Big Five," its implementation as well as its extension in IBM Watson. Successively, in Section 2, we introduce the data used, the method applied and the results obtained in our analysis. In Section 3, we discuss some useful applications and extensions for literature scholars. Finally, we conclude by presenting some considerations for future research.

IBM Watson and the Big Five personality insight

Cognitive computing originated in the early 60s, however, it has been improved dramatically in recent years, achieving significant success through the launch of IBM Watson in 2010. IBM Watson simulates human cognitive systems by implementing advanced natural language processing, information retrieval, knowledge representation, automated reasoning, and machine learning technologies to the field of open domain question answering (Ferrucci et al., 2010; Ferrucci, 2012). Among human cognitive activities, one of the most employed is the capability to understand and forecast other people's personalities. As described in the Big Five theory (Norman, 1963) formulated by scholars in psychology, each individual presents a different aptitude in identifying characteristic patterns of thinking, feeling and behaving in others. The Big Five theory is the main theory on which IBM Watson has been built; it distinguishes between five broad dimensions underlying an individual's personality, namely openness, conscientiousness, extroversion, agreeableness, and neuroticism.

- Openness mirrors the level of scholarly interest, imagination and an inclination for oddities.
- Conscientiousness is the propensity to be reliable, to show self-control and act obediently.
- Extroversion comprises vitality, positive feelings, confidence, amiability and the propensity to look for incitement in the organization of others.
- Agreeableness is the inclination to be merciful and agreeable as opposed to suspicious and adversarial towards others.
- Neuroticism is the propensity to encounter obnoxious feelings, for example, outrage, uneasiness, wretchedness, and powerlessness.

Next to the Big Five personalities, IBM Watson takes into account the concept of "Needs," which are described by the literature as universal needs shared by all human beings (Ford, 2005; Kotler and Armstrong, 2010). Along with Big Five and Needs, IBM Watson takes into account the psychological concepts of Values which are defined as "desirable, trans-situational goals, varying in importance, that serve as guiding principles in people's lives" (Schwartz, 2006). As mentioned on the IBM Watson website[1] "Schwartz summarizes five features that are common to all values: (1) values are beliefs; (2) values are a motivational construct; (3) values transcend specific actions and situations; (4) values guide the selection or evaluation of actions, policies, people, and events; and (5) values vary by relative importance and can be ranked accordingly."

Method and Results

Our objective consists in comparing the personality of the main characters of three Shakespeare's tragedies in a positive versus negative sentimental context. To establish these contexts we use the work of Nalisnik et al. (Nalisnick and Baird, 2013) on sentiment analysis to divide the characters of the play into two groups. The first group is composed of those characters towards which the main character expresses mainly positive sentiments; the second group comprises those characters towards which the main character expresses mainly negative sentiments. In practice, we extracted each instance of continuous speech from the plays[2]. The groups divisions take in account the sentiment valence and the minima of sings that the IBM Personality Insights needs to produce significant results and then assumed that each speech act by one speaker was directed towards the character that spoke immediately before him. We used this assumption to replicate Nalisnik's data but, as he pointed out in his paper, "This assumption does not always hold; it is not uncommon to find a scene in which two characters are expressing feelings about someone offstage.". We retrieved the sentiment valences for each main character[3], reported in the following tables:

Hamlet's Sentiment		**Othello's Sentiment**		**Macbeth's Sentiment**	
Valence Sum		**Valence Sum**		**Valence Sum**	
Guildenstern	31	Iago	71	Murderer 1	22
Polonius	25	Cassio	38	Banquo	16
Gertrude	24	Brabantio	27	Duncan	8
Horatio	12	Duke of Venice	24	Angus:	7
Ghost	8	Montano	7	Macduff	5
Marcellus	7	Desdemona	-1	Which 3	5
Osric	7	Lodovico	-4	Which 1	1
Bernardo	2	Emilia	-10	Young Siward	-4
Laertes	-10			Lennox	-11
Phelia	-5			Seyton	-20
Rosencrantz	-12			Lady Macbeth	-39
Claudius	-27				

Table 1: Tables representing the sentiment valence sum, and the used groups for each main character. Positive scores stand for a character's positive attitude towards others and negative scores stand for a negative attitude

	Positive	**Negative**
Hamlet	6655	3455
Othello	3424	2407
Macbeth	1293	1852

Table 2: Word counts per characters of positive and negative sentimental context

By assembling all continuous speech sequences directed to the characters in the different groups, we created two text groups: the first group contains all lines of the main character towards the others when expressing positive sentiments, the second group includes those speech sequences characterized by a negative sentimental connotation. We performed this task for three main characters, more specifically Hamlet, Othello and Macbeth. In Table 2, the positive and negative word counts for these text assemblies are plotted.

Finally, we used IBM's personality insight service to create personality profiles for each of the main characters based on the positive and negative texts. Figures 1, 2 and 3 represent the distribution of needs and values with respect to the positive or negative sentimental valence. Points on line represent independence with respect to the latter. For needs and values falling below this line, the distribution is characteristic for the expression of negative sentiments. We also created personality graphs with respect to the groups. Although they are calculated as percent of a given facet (as indicated by the x axis), the facets vary between low and high values. Thus, the bars in the graph begin at 50% and can either be low (to the left) or high. A score of 50% signifies that the facet is balanced.

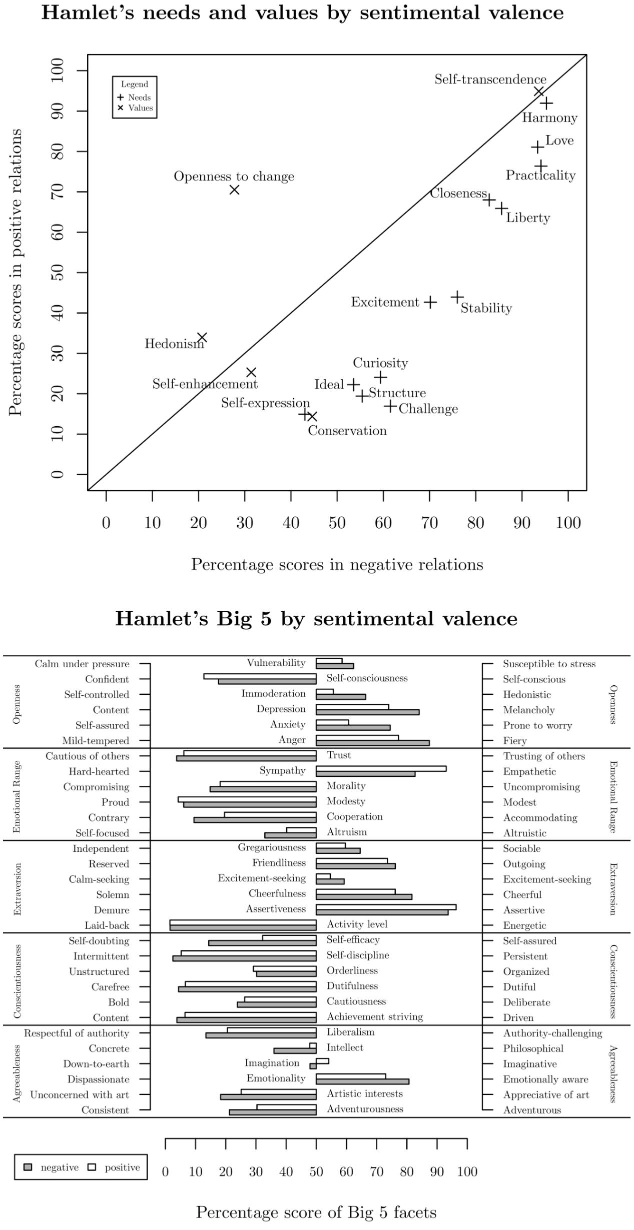


Figure 1: Hamlet's scores for Needs, Values and the Big Five categories

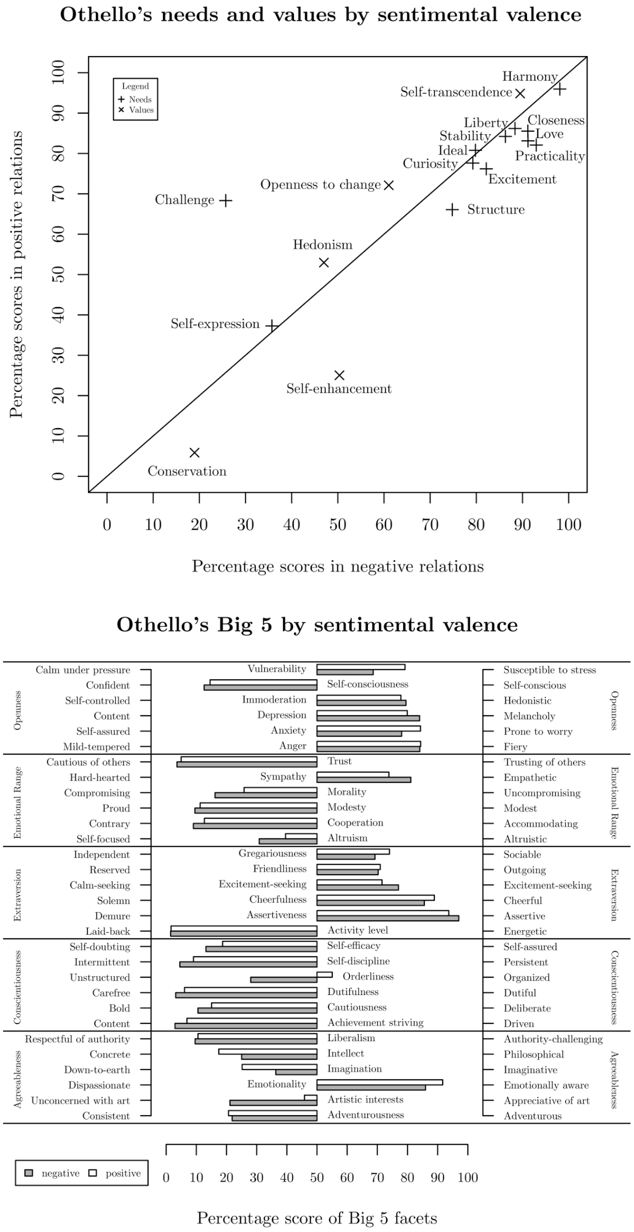


Figure 2: Othello's scores for Needs, Values and the Big Five categories

Some Insights and Discussion

A quantitative approach to character clearly generates a comparatively large amount of linguistic data. The question is, how are we to use this data? What can such an approach offer Shakespeare studies and the humanities more broadly? How does the data we have generated contribute to the critical conversation in a sub-field like character criticism (Hazlitt, 1845; Bradley, 1992; Desmet, 1992; Yachnin and Slights, 2009) which has such a long and distinguished history. There are two answers to this question. First, our analysis shows in a more concrete and detailed way than ever before, the close relationship between character and language, something easy to forget in the context of a representational practice like theatre which is so dependent on non-linguistic features, such as gesture, costume, and stage properties. Playgoers, however, do not just see character; they also hear it. And in the modern humanities classroom, they read it. Accordingly, words play a significant role is crafting what we would now call the "personalities" of Shakespeare's stage. Our approach offers a new means of isolating and analyzing these verbal features of character. The second way in which our work contributes to Shakespeare studies has to do with something our data does not tell us. To understand what we mean by this, consider for a moment why it is that the words associated with some characters generate a personality profile that anyone familiar with the plays knows does not quite fit? Why, for example, does Hamlet have a verbal data set that makes him seem much more serious and honor-driven than he actually is It is because the technology we are using is not capable of accounting for context and therefore cannot detect things like irony and wordplay, two things that are extremely important components of the way language articulates character and personality. While this may be viewed as a methodological weakness from the perspective of information technology, it is a great strength from the perspective of the humanities. For it is precisely at those moments that the data fails to deliver coherent results that we are forced to ask compelling questions about language and art: why do words that seem to mean one thing have the opposite effect on stage? What is the relationship between the literal and implied meanings of words in the constitution of dramatic character? In the end, this facet of our contribution to Shakespeare studies illustrates something important not just about our project specifically, but also about the digital humanities generally: the value of applying computational technology to literary texts lies not in the promise of "better data" or irrefutable "facts," but rather in the way such technology returns us again and again to the fundamental humanist questions that help us understand how literature and art work[4].

Future Research

In this paper we introduced novel techniques based on cognitive computing to get an understanding of characters in Shakespeare's plays. This research explores newway of computer-assisted methods for the investigation of literature. Nonetheless, some technical issues need to be overcome in order to improve the quality of this new methodology. First, an improved methodology to outline the sentiment polarity needs to be developed. Second, the low representativity of IBM's personality model needs to be enhanced in order to catch literary phenomena in sources such as Twitter, Wikipedia, and other such corpora. Thus, recreating a personality insight service based on a literary work corpus would surly enhance the results of our method.

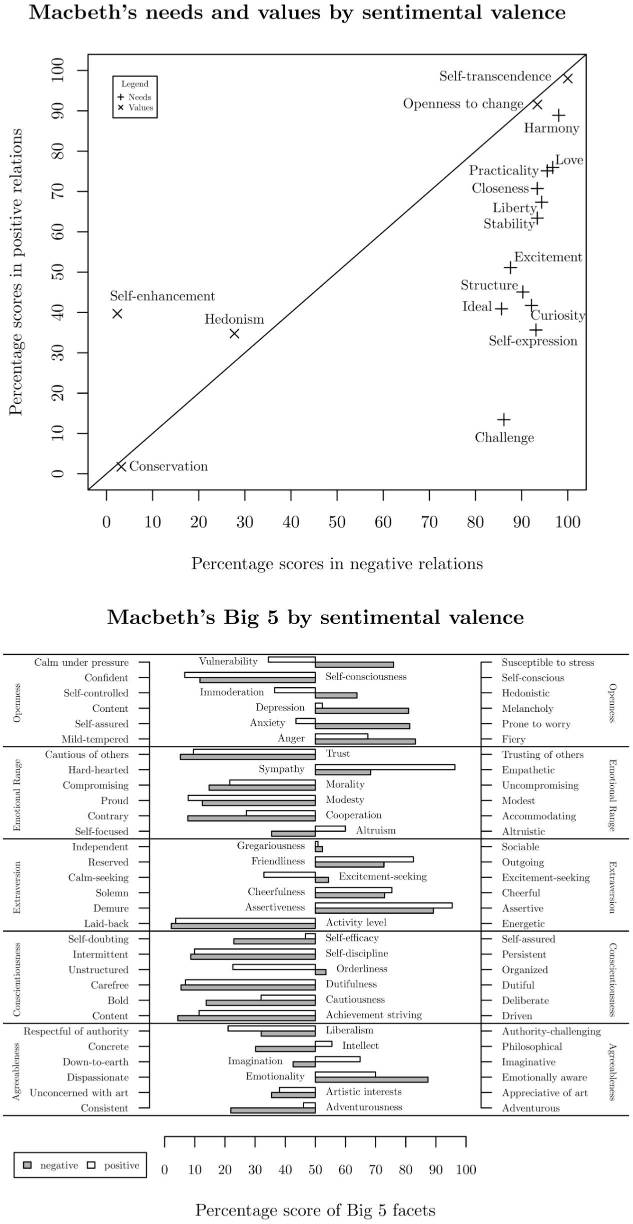


Figure 3: MacBeth's scores for Needs, Values and the Big Five categories

Bibliography

Bradley, A. C. (1992). *Shakespearean Tragedy: Lectures on Hamlet, Othello, King Lear, Macbeth*. 3rd ed. New York: St. Martinis Press.

Desmet, C. (1992). *Reading Shakespeare's Characters: Rhetoric, Ethics, and Identity*. Amherst: Univ of Massachusetts Press.

Ferrucci, D. A. (2012). Introduction to 'This is Watson'. *IBM Journal of Research and Development*, **56**(3): 1 doi:10.1147/JRD.2012.2184356.

Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., et al. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, **31**(3): 59–79.

Ford, J. K. (2005). *Brands Laid Bare: Using Market Research for*

Evidence-Based Brand Management. Wiley https://books.google.it/books?id=oQ3OClLh6UsC.

Gill, A., Brockmann, C. and Oberlander, J. (2012). Perceptions of Alignment and Personality in Generated Dialogue. *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*. Utica, IL: Association for Computational Linguistics, pp. 40–48 http://www.aclweb.org/anthology/W12-1508.

Hazlitt, W. (1845). *Characters of Shakespeare's Plays*. Boston: Wiley and Putnam.

Konstantopoulos, S. (2010). An Embodied Dialogue System with Personality and Emotions. *Proceedings of the 2010 Workshop on Companionable Dialogue Systems*. Uppsala, Sweden: Association for Computational Linguistics, pp. 31–36 http://www.aclweb.org/anthology/W10-2706.

Kotler, P. and Armstrong, G. M. (2010). *Principles of Marketing*. (PRINCIPLES OF MARKETING). Prentice Hall https://books.google.it/books?id=5HkrAQAAMAAJ.

Mairesse, F. and Walker, M. (2006). Automatic Recognition of Personality in Conversation. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York City, USA: Association for Computational Linguistics, pp. 85–88 http://www.aclweb.org/anthology/N/N06/N06-2022.

Mairesse, F. and Walker, M. (2007). PERSONAGE: Personality Generation for Dialogue. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 496–503 http://www.aclweb.org/anthology/P07-1063.

Nalisnick, E. T. and Baird, H. S. (2013). *Character-to-character sentiment analysis in Shakespeare's Plays*.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings. *J Abnorm Soc Psychol*, **66**: 574–83.

Schwartz, S. H. (2006). Basic human values: Theory, measurement, and applications. *Revue Française de Sociologie*, **47**(4): 249–88.

Yachnin, P. and Slights, J. (2009). *Shakespeare and Character: Theory, History, Performance and Theatrical Persons*. Basingstoke: Palgrave Macmillan.

Notes

[1] http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/doc/personality-insights/science.shtmlaccessed on October 25 2015

[2] XML versions provided by Jon Bosan: http://www.ibiblio.org/xml/examples/shakespeare/accessed on October 15 2015

[3] The tables provided by Nalisnik in http://www.ics.uci.edu/~enalisni/ShakespeareExplorer.htmlaccessed on October 15 2015

[4] This idea has been advanced influentially in the work of Jonathan Hope and Michael Witmore. See their blog, Wine Dark Sea http://winedarksea.org.

Visualisation Strategies for Comparing Political Ideas with the ORATIO Platform

Tommaso Elli
tommaso.elli@gmail.com
Politecnico di Milano

Giovanni Moretti
moretti@fbk.eu
Fondazione Bruno Kessler, Italy

Rachele Sprugnoli
sprugnoli@fbk.eu
Fondazione Bruno Kessler, Italy; University of Trento

Michele Mauri
michele.mauri@polimi.it
Politecnico di Milano

Giorgio Uboldi
giorgio.uboldi@gmail.com
Politecnico di Milano

Sara Tonelli
satonelli@fbk.eu
Fondazione Bruno Kessler, Italy

Paolo Ciuccarelli
paolo.ciuccarelli@gmail.com
Politecnico di Milano

Introduction

Data visualisation has become one of the most relevant DH topics, due to the advent of Big Data in Humanities research practices, and to the need to make complex statistical analyses accessible to users without a technical background. Although several visualisation libraries, such as d3.js, are now freely available online and are relatively easy to use, it is still a challenging task to provide simple and effective interface design, avoiding both over-complex and over-simplified solutions. When the data to be displayed have undergone complex processing, for instance automated text analysis, it is of paramount importance to preserve all the information conveyed by such analyses, while making it understandable to the users.

In this work, we present a collaboration between communication design and natural language processing (NLP) researchers, devising effective strategies to display different aspects of the semantic content of texts. The outcome of the collaboration is the ORATIO platform, specifically developed to *compare* different points of view automatically extracted from text. The most challenging tasks, indeed, concerned the visualisation and the exploration of differences and overlaps detected through automated text processing.

Use case

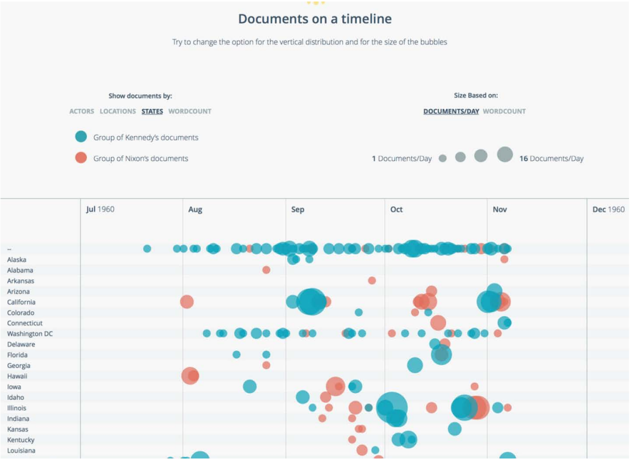
Our use case concerns the comparison between Nixon's and Kennedy's speeches uttered during the U.S. presidential campaign in 1960. The corpus consists of 282 documents by Nixon (830,000 tokens) and 598 documents by Kennedy (815,000 tokens)[1]. The overall goal of the project was to track the difference in language and content between the two opponents, and make it available through a platform which makes use of a "generous interface": first providing all the information to the user, and then enabling him to handle the visual model through a number of options and filters (Whitelaw, 2012). Infact, in our setting, researchers are supposed to reshape and reduce the visualizations in order to prove theories or discover new interesting aspects related to the processed text. The proposed navigation pattern complies with the paradigm "Overview first, zoom and filter, details on demand" (Heer and Shneiderman, 2012).

Other existing approaches do not start from an overview, but from an empty window, where the user can build up a personal view, while investigating the relationships inside the data. We rely on such approaches in order to design the last visual model of the platform (Fig. 6), while the others take from the first one, starting from an overview.

ORATIO Description

To cope with corpora richness, a multiple view approach has been adopted (Mauri, Pini, Ciminieri and Ciuccarelli, 2013): rather than providing a single view, with all the information, five different perspectives have been identified, each exploring a different piece of information in a comparative way. The first view is the *Summary*, whose goal is to provide the user with a general overview of the two corpora, including geographical, temporal and size information. Each corpus is associated with an imagine and a color (blue for Kennedy, red for Nixon), which remain consistent across all the platform views. Under *Summary*, users can see how speeches are distributed on a map (according to the place where the talk was given, included in the metadata), on a timeline (based on day of the speech in the metadata), and what linguistic features characterise each corpus (i.e. number of speeches, average words in a document and total number of words). For instance, in Figure 1 a compact representation of three corpus dimensions is given: the x-axis represents the timeline, the y-axis includes the list of cities where the speeches were given, and the dimension of the bubbles corresponds to the number of speeches uttered in a certain place at a certain time point.

The visualisation shows, for instance, that Nixon pledged to visit all the 50 States, while Kennedy did not held any speech in some States that were less critical to the victory of the elections (e.g. Hawaii or Vermont). Another interesting aspect of the electoral campaign emerging from this view is that, despite having visited less States, Kennedy was more active than Nixon: he stopped in a higher number of cities (239 cities overall, against the 172 cities visited by Nixon), and had about twice as many speeches, press releases, statements and remarks as his opponent (about 550 for Kennedy and 260 for Nixon). This is highlighted by the prominence of blue over red bubbles.



Documents on a timeline

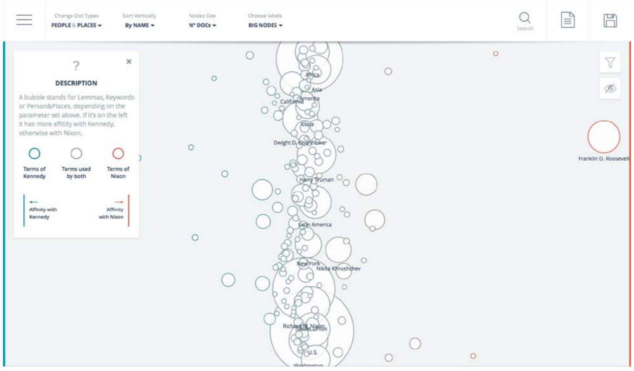
Try to change the option for the vertical distribution and for the size of the bubbles

Show documents by: ACTORS LOCATIONS STATES WORDCOUNT

Size Based on: DOCUMENTS/DAY WORDCOUNT

Fig.1: Summary view of the two speech corpora

The second view, called *Affinity*, targets the need to understand the relevance of topics in the political debate and the presence of important differences between the two candidates. In this view, specific word classes such as verbs, keywords or persons' names are displayed as circles, whose size is proportional to the number of occurrences in text. The more the terms occur in both corpora, the more they are displayed towards the center of the window. If they occur prevalently (or only) in Kennedy's or Nixon's speeches, they are displayed towards the left or the right side of the window, respectively (Fig. 2).



Change Dot Types: PEOPLE & PLACES

Sort Vertically: By NAME

Nodes Size: N° DOCs

Choose labels: BIG NODES

DESCRIPTION

A bubble stands for Lemmas, Keywords or Person&Places, depending on the parameter set above. If it's on the left it has more affitity with Kennedy, otherwise with Nixon.

Fig. 2: *Affinity* view showing the most relevant personal entities discovered in the corpora.

The third view, displaying *People*, gives a network-based representation of the people automatically recognized in the corpora by a Named Entity Recogniser (Finkel et al., 2005). If two or more people are mentioned within the same sentence, they are linked in a spatialized graph. As with the other views, users are then able to filter out elements from the visualization, in order to discover new patterns (Fig. 3a). In our specific use case, filters and other selection strategies are really useful, since the complete network is very large and difficult to read at a glance (Fig. 3b).

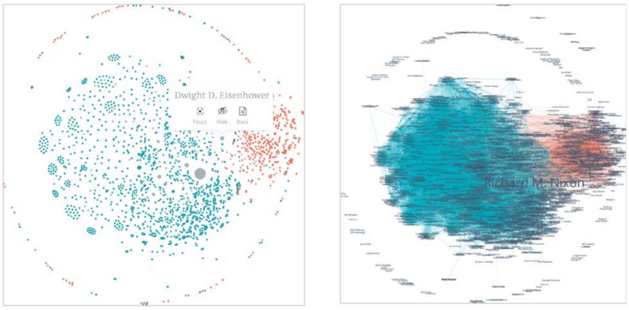


Fig. 3a: *People* view after filtering Fig. 3b: The default network in *People* view

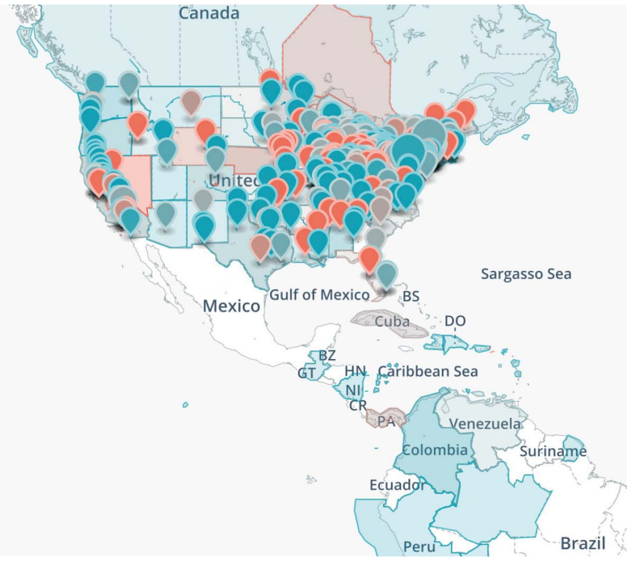
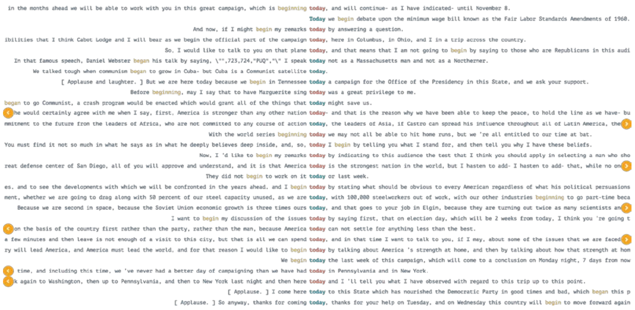


Fig. 4: *Places* view with visited places (marked with cursor) and mentioned places (colored)

The *Places* view provides a comprehensive visualisation of the geographical information contained in the two corpora. It displays the metadata about the place where the speeches were uttered together with the GPEs mentioned in the speeches, automatically extracted with the same Named Entity Recogniser used for *Persons*. These two pieces of information are usually displayed separately, since the most widely used visualisation strategies based on heatmaps would not allow to distinguish them. However, we devised a solution where both can appear on the same map, while being easily distinguishable: the locations where a speech was uttered are marked with a cursor, while the mentioned places are highlighted on the map as colored areas. The comparison shows that Kennedy devoted more attention to specific areas outside US, while Nixon was more concerned with domestic policy. For instance, Kennedy mentioned several times places in Latin America, since one of the key themes of his campaign was the "Good Neighbor" policy, a topic not covered by Nixon.

The last view, named *Concordances*, is inspired by linguistic research and recalls the family of concordancer tools (see for instance Kehoe and Renouf, 2002). In contrast with the previous models, this functionality takes a different approach, since there is no overview and the user is supposed to create a representation in order to answer questions and prove hypotheses. Specifically, a user can look for a particular keyword or concept and see all the sentences where it appears, typographically aligned to ease readability. In a second step, other important terms close by the given concept can be displayed as well (Fig. 5).



```
 in the months ahead we will be able to work with you in this great campaign, which is beginning today, and will continue- as I have indicated- until November 8.
                                                                                                  Today we begin debate upon the minimum wage bill known as the Fair Labor Standards Amendments of 1960.
                                                                  And now, if I might begin my remarks today by answering a question.
ibilities that I think Cabot Lodge and I will bear as we begin the official part of the campaign today, here in Columbus, in Ohio, and I in a trip across the country.
                                                          So, I would like to talk to you on that plane today, and that means that I am not going to begin by saying to those who are Republicans in this audi
   In that famous speech, Daniel Webster began his talk by saying, \"",723,724,"PUQ","\" I speak today not as a Massachusetts man and not as a Northerner.
         We talked tough when communism began to grow in Cuba- but Cuba is a Communist satellite today.
                     [ Applause and laughter. ] But we are here today because we begin in Tennessee today a campaign for the Office of the Presidency in this State, and we ask your support.
                                               Before beginning, may I say that to have Marguerite sing today was a great privilege to me.
began to go Communist, a crash program would be enacted which would grant all of the things that today might save us.
he would certainly agree with me when I say, first, America is stronger than any other nation today- and that is the reason why we have been able to keep the peace, to hold the line as we have- bu
mmitment to the future fron the leaders of Africa, who are not committed to any course of action today, the leaders of Asia, if Castro can spread his influence throughout all of Latin America, the
                                                                    With the world series beginning today we may not all be able to hit home runs, but we 're all entitled to our time at bat.
You must find it not so much in what he says as in what he deeply believes deep inside, and, so, today I begin by telling you what I stand for, and then tell you why I have these beliefs.
                                                                   Now, I 'd like to begin my remarks today by indicating to this audience the test that I think you should apply in selecting a man who sho
reat defense center of San Diego, all of you will approve and understand, and it is that America today is the strongest nation in the world, but I hasten to add- I hasten to add- that, while no on
                                                                    They did not begin to work on it today or last week.
es. and to see the developments with which we will be confronted in the years ahead. and I begin today by stating what should be obvious to every American regardless of what his political persuasions
ment, whether we are going to drag along with 50 percent of our steel capacity unused, as we are today, with 100,000 steelworkers out of work, with our other industries beginning to go part-time beca
   Because we are second in space, because the Soviet Union economic growth is three times ours today, and that goes to your job in Elgin, because they are turning out twice as many scientists an
                                                           I want to begin my discussion of the issues today by saying first, that on election day, which will be 2 weeks from today, I think you 're going t
on the basis of the country first rather than the party, rather than the man, because America today can not settle for anything less than the best.
a few minutes and then leave is not enough of a visit to this city, but that is all we can spend today, and in that time I want to talk to you, if I may, about some of the issues that we are faced
ry will lead America, and America must lead the world, and for that reason I would like to begin today by talking about America 's strength at home, and then by talking about how that strength at hom
                                                                                            We begin today the last week of this campaign, which will come to a conclusion on Monday night, 7 days from now
 time, and including this time, we 've never had a better day of campaigning than we have had today in Pennsylvania and in New York.
k again to Washington, then up to Pennsylvania, and then to New York last night and then here today and I 'll tell you what I have observed with regard to this trip up to this point.
                                                                           [ Applause. ] I come here today to this State which has nourished the Democratic Party in good times and bad, which began this p
                                                           [ Applause. ] So anyway, thanks for coming today, thanks for your help on Tuesday, and on Wednesday this country will begin to move forward again
```

Fig 5: the *Concordances* view, displaying the use of "today", compared with the presence of the term "begin".

Conclusions

We presented the ORATIO platform, specifically developed to compare the content of two different corpora in the political domain. The work is the outcome of a collaboration between researchers in Communication Design and Natural Language Processing applied to Digital Humanities. Although NLP allows to process and extract information from large corpora with minimal efforts, it has drawbacks, which are then inherited by the presented platform. For instance, persons' nodes (Fig. 3) need to be disambiguated in order to merge nodes representing co-referring mentions (e.g. "J. F. Kennedy" and "Jack Kennedy"). Also geo-political entities (Fig. 4) require disambiguation and geo-referencing. This was performed completely automatically, but errors are possible, and this kind of visualisation makes it even more straightforward to spot them.

In order to address these issues, possible solutions could be to *1)* give users the possibility to inspect the content of the documents containing displayed information (from *distant* to *close* reading), and then *2)* give them the possibility to manually correct the displayed information (e.g. drag and drop some elements in the space, delete nodes, etc.). The development of new interfaces enabling such human intervention would be very important and represents the future direction of our research.

Bibliography

Heer, J. and Shneiderman, B. (2012). Interactive Dynamics for Visual Analysis. A taxonomy of tools that support the fluent and flexible use of visualizations. In *Queue*, **10**(2).

Finkel, J.-R., Grenager, T. and Manning Ch. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-70.

Kehoe, A. and A. Renouf. (2002). WebCorp: Applying the Web to Linguistics and Linguistics to the Web. In *Proceedings of WWW2002 Conference*, Honolulu, Hawaii.

Mauri, M., Pini, A., Ciminieri, D. and Ciuccarelli, P. (2013). Weaving data, slicing views. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI - CHItaly '13*. New York, USA: ACM Press, pp. 1–8. doi:10.1145/2499149.2499159

Whitelaw, M. (2012). Towards Generous Interfaces for Archival Collections. In *International Council on Archives Congress*. Retrieved from http://www.ica2012.com/files/data/Full papers upload/ica12Final00423.pdf

Notes

[1] From http://www.presidency.ucsb.edu/1960_election.php

Arduino Circuits and Javanese Puppets: 'Re-materializing' Digital Archives through Tangible Interfaces

Miguel Escobar Varela
m.escobar@nus.edu.sg
National University of Singapore, Singapore

The creation and analysis of digital collections is a key aspect of the digital humanities. Deciding what to archive, how to organize the documents, and how to present them to the public are never easy questions. But in the case of theatre and performance archives, these questions are especially complex since theatre performances are necessarily transient and they depend on the embodied co-presence of actors and spectators. In order to document the embodied and ephemeral nature of theatre performances, digital archives often try to include a wide array of documents (video recordings, photographs, motion capture data, 3d models of theatrical spaces and technical scripts). The makers of digital theatre archives are at pains to contextualize their materials and account for aspects that cannot be easily documented and transmitted. Can interaction design help communicate these aspects? Is it possible to imagine and construct interfaces that can communicate cultural context, transcending the limitations of a computer screen, mouse and keyboard?

To investigate these questions within the context of an Indonesian performance archive, I developed a Tangible User Interface (TUI) using open source hardware (Arduino microcontrollers and an array of sensors) and culturally-coded objects that are important for the performances in the archive. This interface was developed as an interactive artwork for educational museums and as a thought experiment on how tangible, culturally-specific interfaces can constitute instances of academic research outputs in the digital humanities.

The archive in question is the Contemporary Wayang Archive (CWA, http://cwa-web.org), a collection of digital recordings and metadata pertaining to Javanese wayang kulit (shadow puppetry), which I have been developing with my collaborators at the National University of Singapore since 2012. Wayang kulit is the oldest theatre form of Indonesia and one of the most important theatre traditions of Southeast Asia. It consists of a form of puppetry where a dalang (narrator-puppeteer) singlehandedly moves all the puppets, speaks all the character parts, jokes with the audience and directs the musicians. He is at the same time a puppeteer, a storyteller, an orchestra conductor and a stand-up comedian. In order to communicate with the audience and the orchestra, he uses different objects in order to control the progression of the story. For this project, I wired different sensors into these objects in order to develop a TUI that recreates the spatial setup of a conventional wayang performance and three key material components of this setup:

- A kerlir or screen where videos from the CWA are projected.
- The kayon. This puppet is shaped like a leaf and it has different narrative functions. In between scenes, the dalang rotates it around its axis and places it on a banana log at a specific angle (either 45^0, 90^0 or 135^0) to indicate the progression of the story. This visual cue is important for audience members that don't watch the entire show. A conventional performance lasts eight hours and the audience members often come and go, drifting in and out of attention. Depending on the specific angle of the kayon, knowing audience members can estimate the specific moment in the development of the performance (which is divided into three main acts). For my interface, I used an Inertial Measurement Unit (IMU) and a wireless sensor in order to detect rotation (angular velocity) in the puppet. By rotating an actual puppet, users of this interactive artwork can navigate to a different digital video in the collection.
- The cempala. By hitting this wooden mallet against the puppet chest, the dalang cues the musicians to start and stop the musical accompaniment. For the interactive artwork, I wired the cempala to a Piezoelectric sensor to measure vibration. When the users of the artwork knock the cempala against a box that mimics the puppet chest, they can start and stop the videos and additional contextual information appears on the projection screen.

- The usage of these objects is not exactly the same in this interactive artwork display as it is in an actual wayang performance. But it can invite users to think about the importance of materiality and embodiment for this particular theatre tradition. In such a way, this interactive artwork can be considered a piece of digital scholarship, a research output of the DH research into wayang kulit that complements the forthcoming online version of the archive and other publications that might arise from this research project (see, for example www.wayangkontemporer.com).

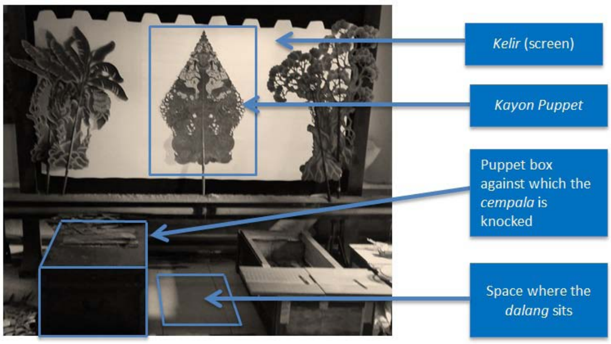

Kelir (screen)
Kayon Puppet
Puppet box against which the *cempala* is knocked
Space where the *dalang* sits

Figure 1. Conventional wayang kulit setup

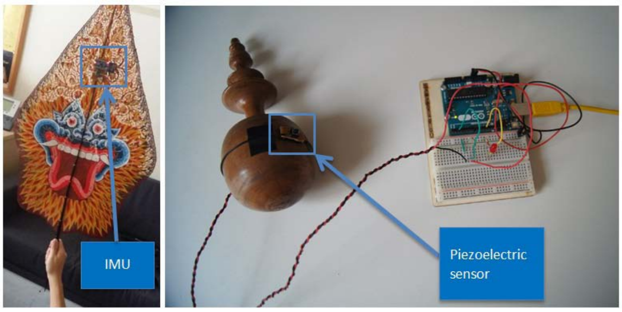


Figure 2. IMU sensor attached to the kayon (left) and piezoelectric sensor attached to the cempala (right)

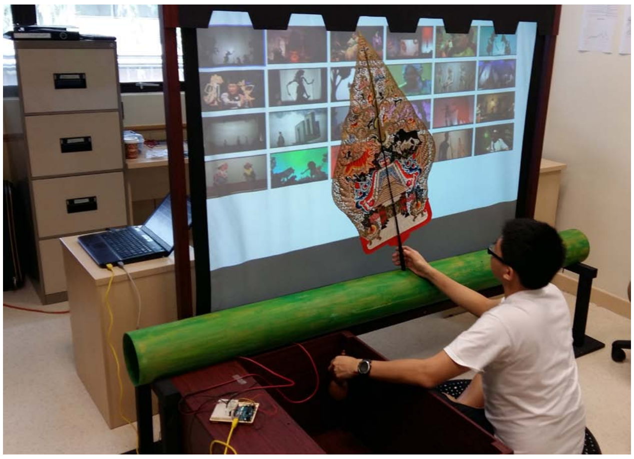


Figure 3. A user interacting with the TUI

- Although this artwork has a very specific origin and function, I hope it will resonate with scholars working in other areas of the digital humanities who are engaged in building and theorizing new artefacts for the communication of academic research. The objective of this short paper is to frame the creation of this artwork within the larger context of the digital humanities, exemplifying modes of scholarship that can emerge at the intersection of cultural knowledge, open hardware and digital technologies.

Bibliography

Fishkin, K. P. (2004). A taxonomy for and analysis of tangible interfaces. *Personal Ubiquitous Computing*, **8**(5): 347–58.

Bonanni, L. et al. (2010). Tangible interfaces for art restoration. *International Journal of Creative Interfaces and Computer Graphics*, **1**(1): 54-66.

Jones, S. et al. (2009). Redefining the Performing Arts Archive. *Archival Science*, **9**(3): 165-71.

Topical Diversification Over Time In The Royal Society Corpus

Peter Fankhauser
fankhauser@ids-mannheim.de
IDS-Mannheim, Germany

Jörg Knappen
j.knappen@mx.uni-saarland.de
Universität des Saarlandes, Germany

Elke Teich
e.teich@mx.uni-saarland.de
Universität des Saarlandes, Germany

Science gradually developed into an established sociocultural domain starting from the mid-17th century onwards. In this process it became increasingly specialized and diversified. Here, we investigate a particular aspect of specialization on the basis of probabilistic topic models. As a corpus we use the Royal Society Corpus (Khamis et al. 2015), which covers the period from 1665 to 1869 and contains 9015 documents[1].

We follow the overall approach of applying topic models to diachronic corpora (Blei and Lafferty 2006, Hall et al. 2008, Griffiths and Steyvers 2004, McFarland et al. 2013, Newman and Block 2006, Yang et al. 2011) to map documents to topics. Probabilistic topic models (Steyvers and Griffiths 2007) have become a popular means to summarize and analyze the content of text corpora. The principle idea is to model the generation of documents with a randomized two-stage process: For every word w_i in a document d select a topic z_k from the document-topic distribution $P(z_k \mid d)$ and then select the word from the topic-word distribution $P(w_i \mid z_k)$. Consequently, the document-word distribution is factored as follows:

$$P(w_i \vee d) = \sum P(w_i \mid z_k)\ P(z_k \mid d).$$

This factorization effectively reduces the dimensionality of the model for documents, improving their interpretability: Whereas $P(w_i \mid d)$ requires one dimension for each distinct word (10s of thousands) per document, $P(z_k \mid d)$ only requires one dimension for each topic (typically in the range of 20-100). Topics are thus not given explicitly for each document, but constitute *latent* variables: A variety of approaches exist to estimate the document-topic and topic-word distributions from the *observable* document-word distributions. We use Gibbs-Sampling as implemented in Mallet (McCallum 2002).

For the preliminary analysis in this paper, we process documents as is, without segmenting them further into pages, only excluding stop words but not performing lemmatization or normalization in order to stay reasonably close to the original source. We experimented with the number of topics ranging between 20 and 30, reporting here results on 24 topics. Cursory analysis of multiple runs with different seeds (Steyvers and Griffiths 2007) shows that the resulting topics are rather stable.

Table 1 displays the top words for the topics with manually assigned labels and their overall percentage of occurrence. We can roughly distinguish four groups of topics; three non-thematic groups and one thematic. The first group comprises topics arising from documents in *Latin* and *French*, some of which are also translated into English. The second group *Formulae* and *Tables* relates to highly formalized modes of information presentation. The third group of topics is also clearly non-thematic but relates to general scientific processes: *Observation* and *Experiment* both contain rather general verbs and adjectives in addition to nouns. *Events* contains words describing remarkable events. *Headmatter* includes formulaic expressions typically occurring at the beginning and end of documents that are letters. All topics in this group are relatively frequent. Finally, the topics in the fourth group (*Geography* through *Chemistry*), consisting mainly of nouns, indeed have a fairly clear thematic interpretation.

Label	**Words**	**%**
Latin	quae quam sed ab sit vero hoc ac sunt esse qui etiam autem pro erit inter quo aut sive	6.4
French	la le les des en du par dans qui il une qu pour ou ce sur ne au je	1.3
Formulae	cos equation sin equal series point equations number line terms form values curie	4.7
Tables	weight water oo oz parts gr grain io grains fat increase weights grs passed urine specific	1.7
Observation	great made make parts found body time small part water nature long good put find	10.4
Experiment	present general subject case results similar nature author state result cases fact	7.3
Events	great time account stone ground house fire letter place miles found side Stones	5.9
Headmatter	years year author society age number time royal life great letter account part letters	5.4
Geography	water sea tide high found river coast north land tides miles height surface great level	3.1
Meteorology	day ditto rain wind cloudy weather fair clear april year days night march july june	3.2
Botany	leaves plant plants tree tab bark folio foliis trees seeds seed flowers species fruit leaf	2.9
Reproduction	cells animal blood fluid eggs membrane found egg part animals ova size Young	3.0
Cells	fibres structure form surface portion cells anterior part section side posteriori	2.7
Paleontology	part bone bones teeth surface upper side lower anterior length posterior tooth large	2.5
Physiology	blood heart muscles part animal nerves vessels left parts stomach bladder body	5.5
Galaxy	distance position stars star obs small hill double equatorial vf diff st magnitudes cebula	1.6
Terrestrial Magn.	observations needle ship magnetic direct force made variation observed north diurnal	2.6
Solar System	sun time observations moon made observed difference observation clock latitude	5.5
Thermodyn.	air water heat temperature experiments tube experiment glass made time Merkury	4.2
Mechanical Eng.	made length weight end diameter iron instrument experiments brass part point Line	4.5
Electromagn.	force electricity current wire action body power direction fluid motion surface effect	3.8

Optics	light rays glass eye red colours spectrum colour surface lines angle white blue object	3.7
Metallurgy	water acid salt grains quantity iron found solution colour substance experiments gold	4.8
Chemistry	acid water solution gas oxygen hydrogen carbonic cent action obtained salt potash	3.4

Table 1: Top words and percentages for topics

To investigate topical trends in the corpus we follow the approach in (Hall et al. 2008), by averaging the document-topic distributions for each year *y*:

$$P(w_i \mid y) = 1/n \sum P(z_k \mid d_j)$$

with *n* the number of documents in a year. *Figure 1* shows a selection of five topics with the most pronounced change over time. Interestingly, some of the major changes occur for non-thematic topics: The topic *Observation* declines sharply from over 30% to less than 1%. The topics *Experiment* and *Formulae* on the other hand increase starting around 1750. This indicates a substantial paradigm shift over time. Indeed, as Gleick (2010) vividly describes, the early stages of the Royal Society were largely devoted to observing and reporting about natural phenomena. The non-thematic topic *Latin* reaches its peak in the early 18th century, and the thematic topics *Cells* and *Chemistry* show a clear increase with the beginning of the 19th century.

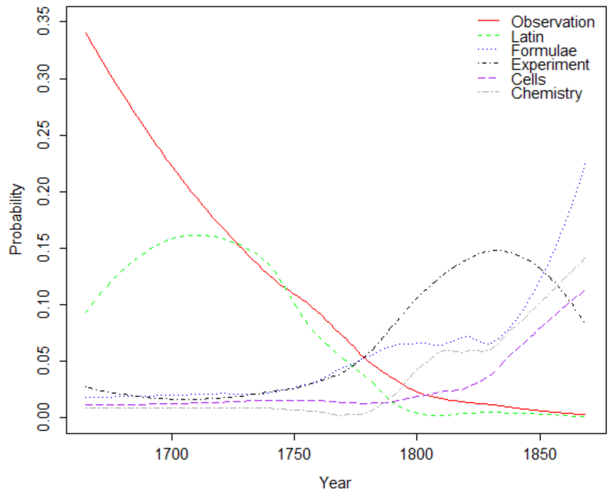


Figure 1: Major topical trends for selected topics

To gain a better understanding about the correlation of topics, we cluster them hierarchically on the basis of the Jensen-Shannon divergence between the topic-document distributions:

$$P(d \vee z) = P(z \mid d) \,/ \sum P(z \mid d_j)$$

Topics that typically co-occur in documents have similar topic-document distributions, and thus will be placed close in the tree.

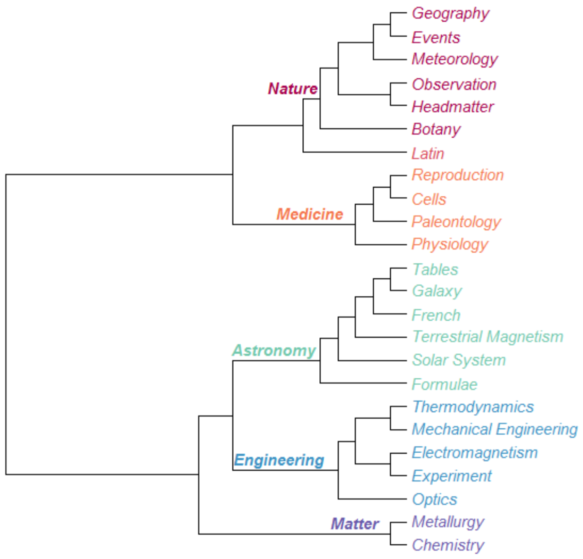


Figure 2: Hierarchical clustering of topics by their topic-document distribution

The resulting tree in *Figure 2* indeed identifies meaningful subgroups. Cutting the tree into six groups - *Nature, Latin, Medicine, Astronomy, Engineering, Matter* - allows us to investigate the overall topic distribution over time (*Figure 3* with *Latin* left out):

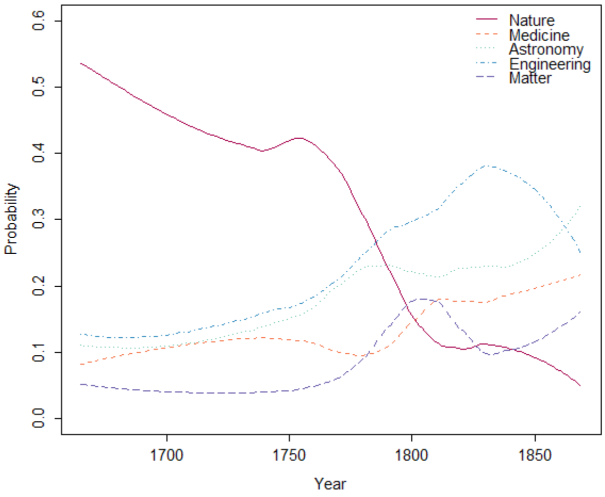


Figure 3: Distribution of topic groups over time

The topic group *Nature* comprising reports of all kinds of natural phenomena (Gleick 2010) clearly decreases over time, which is partially to be attributed to the strong decrease of the topic *Observation* in this group. The topic groups *Medicine* and *Astronomy* increase over time, whereas the topic groups *Engineering* and *Matter* also generally increase but with some intermediate peaks. Similar to the overall trends at the level of individual topics (*Figure 1*), the biggest overall change occurs in the 2nd half of the 18th century.

Looking at the individual trends together, *Figure 3* clearly indicates topical diversification: Until around 1770, the dominance of the topic group *Nature* leads to a highly skewed distribution of topic groups, whereas after 1770 topic groups are distributed much more evenly. The amount of skew can be characterized by the Shannon-Entropy:

$$P(d_y) = -\sum P(z_k \mid y) \log_2 P(z_k \mid y)$$

of the year-topic distributions $P(z_k \mid y)$ (Hall et al. 2010), with highly skewed distributions having low entropy. Indeed as can be seen in *Figure 4 (left)*, the entropy (*ent*) increases fairly consistently during the 18th century and levels out during the 19th century, reflecting a general increase of topical diversity over time.

It is interesting to compare this with the mean entropy of the *individual* document-topic distributions (*ment*):

$$H_{\text{mean}}(P_y) = 1/n \sum H(P_d)$$

with n the number of documents d_j in year y. This measure decreases over time, i.e., while the overall topical diversity increases, the individual documents become more specific in terms of their topic distributions.

The difference between the entropy of year-topic distributions and mean entropy of individual document-topic distributions,

$$JS\,(P_y) = H(P_y) - H_{\text{mean}}(P_y)$$

is the Jensen-Shannon divergence, which is usually applied to two distributions, generalized to the n topic distributions of all documents published in year y. The two opposing trends of these quantities lead to a constantly increasing Jensen-Shannon divergence, with a particularly sharp increase between 1750 and 1800. *Figure 4 (right)* depicts similar trends based on the 24 individual topic distributions. At this level, the year-topic entropy (*ent*) shows less of a clear trend, but mean entropy (*ment*) also clearly decreases, and consequently the Jensen-Shannon divergence clearly increases. Thus, at both levels of abstraction we can observe a clear diversification of the topics assigned to the individual documents. This strongly indicates a growing separation of individual scientific disciplines over time.

As an alternative perspective on topical entropy *Table 2* gives examples of authors with more than 20 papers. The first three authors have the lowest entropy. The dominating topics for Cayley and Owen clearly characterize their main theme of work. Conversely, Rev. John Swinton's top topic *Headmatter* (62%) does not really reflect the overall theme of his publications (Orientalism), but rather their style as letters to members of the Royal Society – the dominant form of publication in this period. The second three authors have the highest entropy, their three top topics together amount for less than 50% of their overall topic distribution. However, they do characterize the main line of work of the authors in question fairly well.

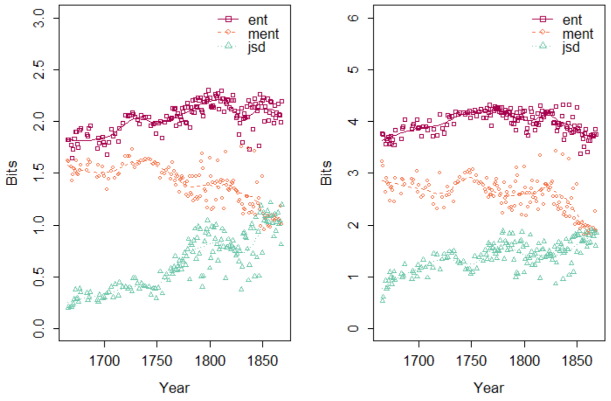


Figure 4: Entropy (ent), mean Entropy (ment), and Jensen-Shannon Divergence for topic groups (left) and individual topics (right)

Author	Papers	Ent	Ment	Jsd	Years	Top Topics
Arthur Cayley	30	1.26	1.12	0.14	1850-1866	Formulae
Richard Owen	26	1.83	1.58	0.25	1843-1869	Paleontology
John Swinton	35	2.50	2.05	0.45	1753-1774	Headmatter
John Davy	58	4.05	3.33	0.72	1800-1856	Experiment, Chemistry, Physiology
William Watson	39	4.03	3.09	0.95	1739-1778	Events, Observation, Botany
Edmond Halley	65	3.93	2.75	1.18	1683-1731	Solar System, Observation, Latin

Table 2: Authors with minimum entropy (top) and maximum entropy (bottom)

In this paper we have analyzed the progression of topics in a corpus of the Royal Society of London. Our main result is the observation that the overall mixture of topics becomes more diverse over time, while the topics of individual documents become more specialized. These two opposing trends lead to a topical fragmentation of scientific discourse, which can be quantified by means of the generalized Jensen-Shannon divergence between the topic distributions of individual documents per time period. We are currently working on consolidating our analysis, experimenting with documents segmented into pages, focusing the analysis on different text types, and more carefully evaluating the resulting topic models (McFarland et al. 2013).

Of course, topic models only provide one, rather broad perspective on diversification of domain specific language. We plan to apply our approach also to other levels of linguistic analysis, such as terminology or grammar.

Bibliography

Blei, D. and Lafferty, J.D. (2006). *Dynamic topic models*. ICML.

Gleick, J. (2010). At the Beginning: More Things in Heaven and Earth: Bryson, B. (Ed.), *Seeing Further. The Story of Science and The Royal Society*. Harper Press, pp. 17-36.

Griffiths, T. L. and Steyvers M. (2004). Finding scientific topics. *PNAS*, 101 Suppl **1**:5228–35.

Hall, D., Jurafsky, D., and Manning, C.D. (2008). Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363-71.

Khamis, A., et al. (2015). A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus. *Corpus Linguistics 2015*. Lancaster.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

McFarland, D. A., et al. (2013). Differentiating language usage through topic models. *Poetics*, **41**(6): 607-25.http://dx.doi.org/10.1016/j.poetic.2013.06.004.

Newman, D. J. and Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *J. Am. Soc. Inf. Sci. Technol.*, **57**(6): 753-67. DOI=http://dx.doi.org/10.1002/asi.v57:6

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. Landauer, T., et al.(Eds.), *Handbook of Latent Semantic Analysis*. Hillsdale, NJ: Erlbaum.

Yang, T., Torget, A. J., and Mihalcea, R. (2011). Topic modeling on historical newspapers. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH '11). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 96-104.

Notes

[1] Of these 205 years, 159 years actually contain documents (mean = 56.7, median=36, sd=61.6, min=12, max=444)

Writing Composition in the Close Reading Cycle: Developing The Annotation Studio Idea Space

Kurt Fendt
fendt@mit.edu
Massachusetts Institute of Technology, United States of America

Suzanne Lane
stlane@mit.edu
Massachusetts Institute of Technology, United States of America

Andy Kelleher Stuhl
akstuhl@mit.edu
Massachusetts Institute of Technology, United States of America

Building on the success of Annotation Studio—MIT's collaborative, open source annotation tool—a new application called the Idea Space expands the functionality of the platform to cover the point in the cycle of close reading where annotation turns to writing composition. This new interface allows students to arrange annotations into an outline, which, when exported to a word processor, retains citations of the original text along with full metadata and links to their annotations. This paper discusses the basis for the Idea Space in assessment of student annotation and in conceptions of digitally aided student scholarship. It also describes the development of the Idea Space as a modular application, complementary to Annotation Studio but easily adaptable to other environments, and presents a working prototype of the Idea Space.

The Annotation Studio online application has been used successfully over the past three years as a tool to support student textual annotation and collaborative reading, and its user base has expanded exponentially over the past two years in educational institutions around the world. In addition to facilitating textual annotation and reading-group formation, Annotation Studio has the potential to provide a writing space in which student annotations may be organized and used as the basis for the development of arguments by the individual or by collaborating groups. We are working, in effect, to develop a pedagogy and a tool that will support a seamless integration of the processes of close reading, annotation and writing. The aim of this new tool, the Idea Space, is to help students learn to collect sources, analyze content, form arguments, properly cite sources and confidently manage the writing process.

Recent assessment has shown that deep engagement with texts through forms of close reading and annotation can have a significant impact on students' ability to become better academic writers. A few studies have pointed specifically to the role of collaborative annotation in strengthening and expanding students' annotation practices, their engagement with textual analysis, and their ability to incorporate and analyze sources in their written work. Yet, collaborative annotation, supported by an online social annotation tool, such as HyperStudio's Annotation Studio, is a relatively new pedagogy, and initial adoption of Annotation Studio has focused on its efficacy as a tool to improve students' critical reading. Preliminary experimentation has suggested that the tool can help students visualize verbal patterns in texts; deepen their in-class discussions; anchor their claims with more detailed textual evidence; dig deeper into contextual and cultural meaning; and provide more thoughtful peer review of each other's texts in progress. Other parts of the writing process, such as organizing evidence, analysis, and claims into coherent argument structures, lie just outside the scope of the current implementation of Annotation Studio.

The Annotation Studio team have therefore prototyped a new "Idea Space" extension of Annotation Studio, in which each student's annotations are displayed and organized around key terms. In this space, the student's or group's annotations (including images and other media) may be sorted, arranged and structured to form arguments and provide the basis for developing supporting materials that compose the substance of academic writing. The Idea Space will help students organize citations, comments, and preliminary text to create an outline of an essay which in turn can be fed back into the annotation tool for further revision and for review by writing instructors. The Idea Space will provide both a workshop space for the development of writing and a window into the writing process itself, from which an instructor will be able to engage with the student.

While the Idea Space presents a novel interface, its scholarly and pedagogical basis is as grounded in ancient practices as annotation itself. When Juliet Fleming takes on the question of how we can usefully define "what reading is," she proposes the metaphor of "cutting." Fleming's suggestion—that reading has always been a process of pulling material from a text and adapting it to the reader's purposes—resonates with the way in which many students currently use Annotation Studio: first flagging evidence in the text through the collaborative annotation process, then later drawing on these same pieces of evidence to compose written arguments. Just as digital interfaces help realize an ideal environment for annotation, they also offer myriad ways in which to support and enrich this process of cutting. The ease of sorting, filtering, duplicating and rearranging material in digital settings affords a fast and intuitive way to turn from annotative reading to composition. The capacity of these virtual materials to retain links to one another, keeping track of where and when annotations or edits have been made, can not only relieve reader-authors of some of the work of managing citations, but also give scholars a greatly augmented record of their own process. This record can, persist throughout the scholarly cycle when works exported from the Idea Space are themselves uploaded into Annotation Studio.

While the Idea Space is conceived as adding to Annotation Studio's support for student work, its modular design and implementation enable it to work in combination with other tools and sources of data. The application's first data module retrieves annotations from the Annotation Studio database, whose format conforms to open annotation standards, meaning that the interface could easily extend other annotation environments. A simple design for adding new data modules gives instructors the ability to tailor the application for use with their own combinations of tools and archives. Far more than a visual organizer for annotations, the Idea Space is a tool through which users can combine, juxtapose and adapt any scholarly material in the composition of compelling and richly contextualized writing.

Bibliography

Fleming, J. (2010). Afterword, *Huntington Library Quarterly* 73 (3): 543–52. doi:10.1525/hlq.2010.73.3.543.

Seeing Andalucia's Late Gothic heritage through GIS and Graphs

Patricia Ferreira Lopes
pwanderley@us.es
University of Seville, Spain

Francisco Pinto Puerto
fssp@us.es
University of Seville, Spain

Antonio Jimenez Mavillard
ajimene6@uwo.ca
University of Western Ontario

Juan Luis Suárez
jsuarez@uwo.ca
University of Western Ontario

1. Introduction

This study discusses the methodology used in *The Digital model of Andalusia's Late Gothic Heritage* project to develop new models of heritage interpretation through the application of GIS and Graph visualization to provide new perspectives of Andalusia's heritage by considering social, political, economic and cultural evolutions. Given the special period context considered, late 15th and early 16th centuries, the project used a variety of sources to relate heterogeneous historical data on different subjects in order to create a historical spatial database and to respond questions such as: How was the construction process in Andalusia between 1433 and 1560? What period and region had more constructive activity? How did the opening of the Andalusia´s eastern border modify the dynamism of the territory? What features are more common in each time period? What professionals have worked together on a building or quarry? Is this reflected in the architectural language produced?

Andalucía's territory is the consequence of a huge numbers of different cultures that passed through it - from Roman and Islamic to Christian. During the 15th century, Spain had profound transformations, both political and cultural. The Christian reconquest led to the centralization of the cultural production in major cities generating important flow of new knowledge. Also, the consolidation of the Iberian Peninsula´s borders and the growth of contacts with Europe, and in a near future with America as well, helped to improve the cultural expansion and exchanges between the political leaders, intellectuals, scientists and technicians (García C., 2011).

These transformations are documented by a large number of sources that usually focus only on one particular aspect. In this sense, some sources are about professionals that worked at the quarries, others are about building construction and labor contracts, or about journeys and meetings that the workers had had. On one hand, we are dealing with the data from researchers that studied the biography of a particular professional or studies that have been dedicated to the constructive evolution of a city or a building. On the other hand, we are analyzing historical cartography sources that show different territorial changes and stages over time (kingdoms, dioceses, borders, etc).

Thereby, our goal is to build a spatial-temporal database capable of linking information that at first glance may seem unrelated. Visualizing and relating these attributes through an information system on cultural heritage has steered our work in two directions: the creation of systems built around the entities, and the implementation of analyses to observe and interpret their relationships.

In this study we demonstrate how all different historical information could be organized and structured in two types of digital models - GIS and graphs - that will are applied to our case study and allow us to generate a more comprehensive and flexible understanding of the phenomenon of the late Gothic period as a complex system through a combined knowledge of space, time and actors.

Working with these two tools has provided new perspectives at the Late Gothic heritage, creating new groups and subgroups of entities, and new relationships that could be easily translated to other case studies. We also created different categories - for each technology - designing two models of organizational structure. On the one hand, we use GIS with a spatial approach, the space is the product and simultaneously the producer of a series of relationships (Lefebvre, 2000) whose analysis can be performed using the alphanumeric attributes of each spatial entity or its topology. On the other hand, using graphs we have an abstract approach to visualizing a network of professionals and works over the Late Gothic period. The creation of the two models is because each of these tools requires a specific way of organizing data. While GIS works with a system with fixed relational databases that support SQL "join" operations, using it as the main query language and organizes spatial data in layers and attributes, Graph model uses the NoSQL system (Not Only SQL) and organizes data in nodes and edges (Robinson et al., 2013).

2. GIS model

The process of creating the GIS model will be developed in eight dynamic and interrelated phases: database design schema (figure 1); collection, processing and data selection; data entry and analysis in ArcGIS software; production model views; assessment of the problem; data interconnection; generating queries and reports; system development documentation and dissemination of the model (Ferreira Lopes and Pinto Puerto, 2015). The longest stage of the process is the creation of spatial entities - data entry - around 75% of the time and effort undertaken in the research will be used to collect, treat and create the data so that they can later be analyzed. The big difficulty lies in its accuracy. For certain entities, the maps and information of the 15th and 16th centuries does not allow us to reach an urban scale precision which forced us to work with a territorial scale. That is the case, for example, of the kingdoms, buildings, quarries or paths layers (figure 2).

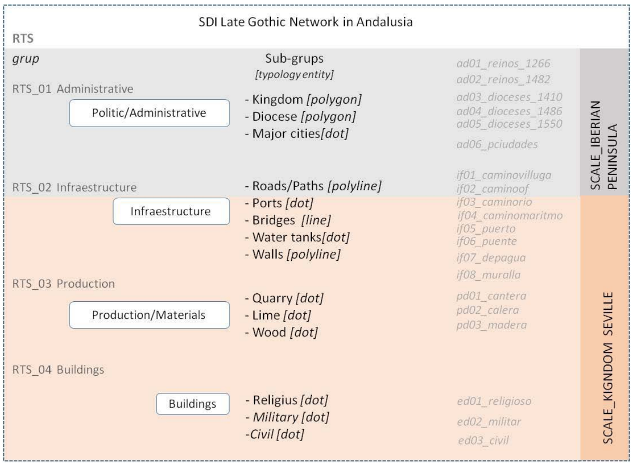
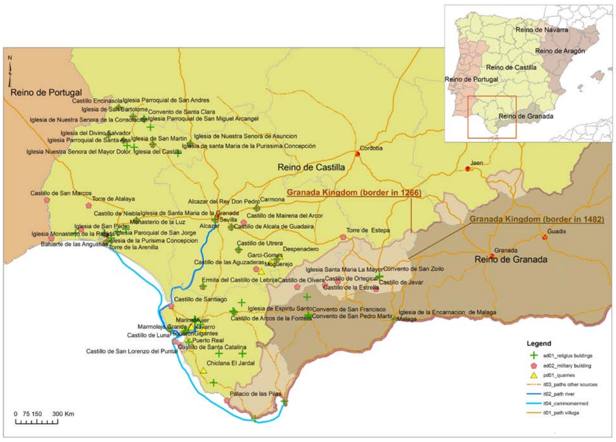


Figure 1: Andalucía's Late Gothic Heritage database schema

Also, some gaps are constantly present in the attributes; in many cases we have incomplete information, which somewhat limited our analysis of certain data. That is the case, for example, of some *"date"* attributes. However, working with GIS allows us to take these gaps temporarily - once it offers an easy way to edit and add new data, which is one of the main advantages that the tool offers. Our SDI are created in the ISO 19100 series standards, Open Geospatial Consortium standards and the recommendations of the guidelines of INSPIRE and LISIGE in order to be extended, edited and viewed by other researchers.



Reino de Navarra
Reino de Aragón
Reino de Castilla
Reino de Portugal
Reino de Granada
Reino de Portugal
Castillo Encinasola
Iglesia Parroquial de San Andres
Iglesia de San Bartolome
Convento de Santa Clara
Iglesia de Nuestra Senora de la Consolación
Iglesia Parroquial de San Miguel Arcangel
Iglesia del Divino Salvador
Iglesia Parroquial de Santa Ana
Iglesia de San Martin
Iglesia de Nuestra Senora de Asuncion
Iglesia de santa Maria de la Purissima Concepción
Iglesia Nuestra Senora del Mayor Dolor.
Iglesia del Castillo
Cordoba
Reino de Castilla
Jaen
Castillo de San Marcos
Torre de Atalaya
Alcazar del Rey Don Pedro
Carmona
Granada Kingdom (border in 1266)
Castillo de Niebla
Iglesia de Santa Maria de la Granada
Sevilla
Castillo de Mairena del Arcor
Monasterio de la Luz
Iglesia de San Pedro
Alcazar
Castillo de Alcala de Guadaira
Granada Kingdom (border in 1482)
Iglesia Monasterio de la Rabida
Iglesia Paroquial de San Jorge
Torre de Estepa
Guadix
Baluarte de las Angustias
Iglesia de la Purisima Concepcion
Torre de la Arenilla
Castillo de Utrera
Despenadero
Granada
Reino de Granada
Garci-Gomes
Castillo de las Aguzaderas
Moguerejo
Iglesia Santa Maria La Mayor
Convento de San Zoilo
Ermita del Castillo de Lebrija
Castillo de Olvera
Castillo de Ortegica
Castillo de Jevar
Castillo de la Estrella
Castillo de Santiago
Iglesia de Espiritu Santo
Convento de San Francisco
Convento de San Pedro Martir
Iglesia de la Encarnacion de Malaga
Malaga
Castillo de Arcos de la Fontera
Marina Mujer
Marmolejo Grande
Navarro
Castillo de Luna
Higueron
Gigantes
Puerto Real
Castillo de Santa Catalina
Castillo de San Lorenzo del Puntal
Chiclana El Jardal
Palacio de las Pilas
Legend
ed01_religius buildings
ed02_military building
pd01_quarries
it03_paths other sources
it02_path river
it04_caminomarmed
it01_path villuga
0 75 150 300 Km
N

Figure 2: Andalucía's Late Gothic Heritage SDI, more than 100 buildings, 4.000km of paths, administrative and dioceses borders (which has changed through time), 14 quarries, and others entities

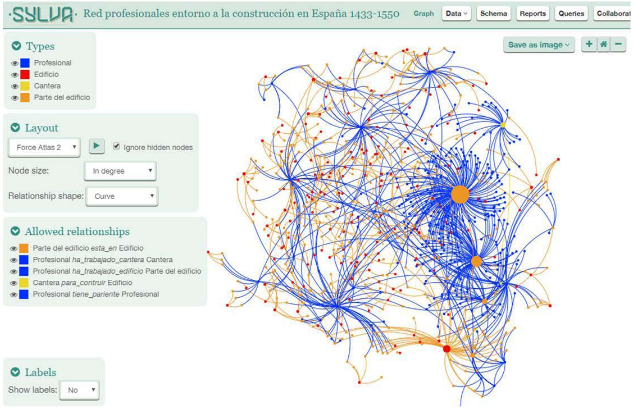
3. Graph Model

The construction process of the Graph Model has been developing in three phases: 1) data collection, 2) scheme creation and 3) queries and analysis. This paper will deal with the proposal and outline gathering held in conjunction with the Culture Plex Lab at the University of Western Ontario in Canada.

Our "starting point" was the Cathedral of Seville. Knowing that the construction of the Cathedral was one of the events most responsible for the flow of knowledge at that time and knowing that a lot of professionals had worked or had some kind of work relation with it, this strategy was more realistic and reachable, initially (figure 03).

To create the Graph model we are using the software Sylva DB (de la Rosa et al., 2013) developed by the CulturePlex Lab, which has allowed us to create a scalable and flexible way to organize, structure, manage, visualize and analyze our mass of data. With all the data in *SylvaDB*, we can see the links between different professionals and works, as well as between quarries and workers, parts of building and professionals, quarries and buildings. This tool has redirected the process in a way that otherwise would have been too taxing in terms of time and computation - we were now not required to create innumerable tables and charts to collect all the information.

After having a clear starting point and already with a certain amount of professionals collected - about 300 workers - the next step was the creation of a schema through which we discern graphic patterns. At this moment, in the Graph model we have 1.000 relationships and 850 nodes (figures 04, 05 and table 1).


·SYLVA· Red profesionales entorno a la construcción en España 1433-1550

Graph | Data | Schema | Reports | Queries | Collaborat

Save as image | + | 🏠 | −

Types
- 🟦 Profesional
- 🟥 Edificio
- 🟨 Cantera
- 🟧 Parte del edificio

Layout
- Force Atlas 2 ▶ ☑ Ignore hidden nodes
- Node size: In degree
- Relationship shape: Curve

Allowed relationships
- 🟧 Parte del edificio *esta_en* Edificio
- 🟦 Profesional *ha_trabajado_cantera* Cantera
- 🟦 Profesional *ha_trabajado_edificio* Parte del edificio
- 🟨 Cantera *para_contruir* Edificio
- 🟦 Profesional *tiene_pariente* Profesional

Labels
- Show labels: No

Figure 3: *Sylva´s* print screen showing the Andalusia´s Late Gothic Network Graph model

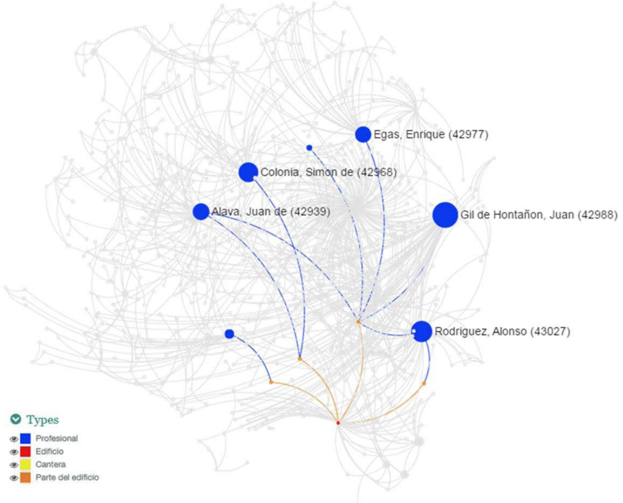


Figure 4: This graph shows the masters builders who had worked in different activities concerning to the cimbor of the Cathedral of Seville

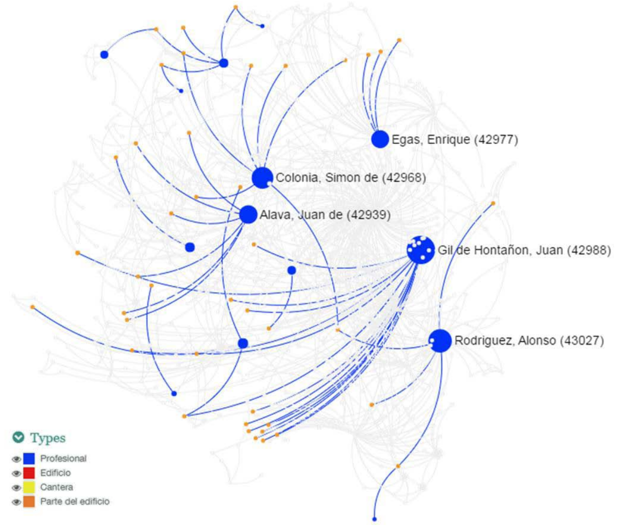
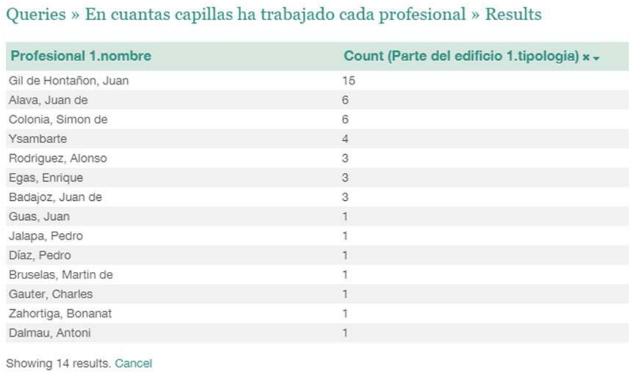


Figure 4: This graph shows the masters builders who had worked in some chapels



Queries » En cuantas capillas ha trabajado cada profesional » Results

Profesional 1.nombre	Count (Parte del edificio 1.tipologia) × ▾
Gil de Hontañon, Juan	15
Alava, Juan de	6
Colonia, Simon de	6
Ysambarte	4
Rodriguez, Alonso	3
Egas, Enrique	3
Badajoz, Juan de	3
Guas, Juan	1
Jalapa, Pedro	1
Díaz, Pedro	1
Bruselas, Martin de	1
Gauter, Charles	1
Zahortiga, Bonanat	1
Dalmau, Antoni	1

Showing 14 results. Cancel

Table 1: This table shows the professionals who worked in different chapels. We can see the importance of some of them - the ones who more chapels worked in

4. Conclusions

Different research methods based on new technologies applied to understanding the same phenomenon provide a greater depth of the problem. In this sense, the use of multiple methods has allowed us to achieve three important aspects: to promote different perspectives on the subject allowing a wider view about the object of study; include a large variety of variables in the study; provide multiple analyses of the same concept, which increases the validity of the research that remains open and upgradeable. Therefore, what we seek is to provide new methods but also new approaches that do not override other traditional systems, but enrich the discussion on the past and its relationship with the inheritance.

Funding

This project is associated with the project and R+D+I (Research, Development and Innovation) *"A Digital Information Model for the Knowledge and Management of Immovable Cultural Heritage"* funded by the Spanish Ministry of Economy and Competitiveness and led by the lecturer Francisco Pinto Puerto.

Bibliography

de la Rosa J., Suárez J. and Sancho Caparrini F. (2013). SylvaDB: A Polyglot and Multi-backend Graph Database Management System. In *Proceedings of the 2nd International Conference on Data Technologies and Applications*, pp. 285-92. http://sylvadb.com. University of Western Ontario, Canada.

Ferreira Lopes, P. and Pinto Puerto, F. (2015). Application of a schema to late gothic heritage: creating a digital model for a spatio-temporal study in Andalusia. In *WIT Transactions on the Built Environment*. UK: Wessex Institute Press, **153**: 29-41.

García Cuetos, M. P. (2011). Raíces del Tardogótico castellano. La arquitectura europea en el contexto del último gótico: ¿una arquitectura paneuropea? In Alonso Ruiz, B. (eds), *La arquitectura tardogótica castellana entre Europa y América*. Madrid: Silex.

Lefebvre, H. (2000). *La producción del espacio*. Madrid: Capitan Swing.

Robinson, I., Webber, J. and Eifrem, E. (2013). *Graph Databases. Information Management*. O´Reilly Media, Inc.

Choosing Words for Stylometric Authorship Attribution: Evaluating Most Distinguishing Words (MDWs) vs. Most Frequent Words (MFWs)

Paul J. Fields
pjfphd@byu.net
Brigham Young University, United States of America

Larry W. Bassist
larrybassist@comcast.net
Brigham Young University, United States of America

Matthew R. Roper
matt_roper@byu.edu
Brigham Young University, United States of America

Introduction

The results of stylometric authorship attribution studies are strongly influenced by four choices:

1. **Candidate Authors** – The choice of candidate authors should be based on the historical context of the texts to be attributed.

2. **Representative Texts** – Representative texts should be chosen that are similar in genre, topic and time frame as the texts to be attributed (Argamon et al, 2003; Stamatatos, 2009).

3. **Analytical Method** – Many analytical methods are available. Burrows' Delta is often considered to be the 'gold standard' to compare other methods (Burrows, 2002).

4. **Distinguishing Features** – One list of features to distinguish among candidate authors can provide greater distinguishing power than another list of features. This paper is about identifying the most distinguishing list.

Grammatical function words are used by all authors, but authors do not use function words in the same way or with the same frequencies. Therefore, different usage frequencies for function words are useful in characterizing an author's writing style. Although the specific function words that are distinguishing among authors vary from study to study, their effectiveness as features to set apart an author's writing style is well established (Mosteller and Wallace, 2007; Holmes, 1998).

Discriminant analysis is a statistical technique to classify objects into known categories based on a set of features about those objects. The technique was developed by Sir Ronald Fisher, a botanist. He illustrated the technique by classifying iris flowers into three species using four features – the length and width of sepals and petals (Fisher, 1936).

The approach is to compute linear combinations of the features that best separate the categories from each other. The most distinguishing combination of features is called the first linear discriminant function (LD1). Additional combinations (LD2, LD3 and so on) are computed that are orthogonal to each other to maximize the separation among categories. After computing the discriminant functions using a training set of data for objects with known classification, the discriminant functions can be used to classify objects of unknown classification into the categories to which they most likely belong.

The discriminant analysis concept is illustrated in figure 1 for a two-category problem and two dimensions. Each ellipse in the graph encircles the items within a category. LD1 shows the direction of greatest separation between the two categories. Discriminant analysis can be extended to classification problems with any number of categories and dimensions.

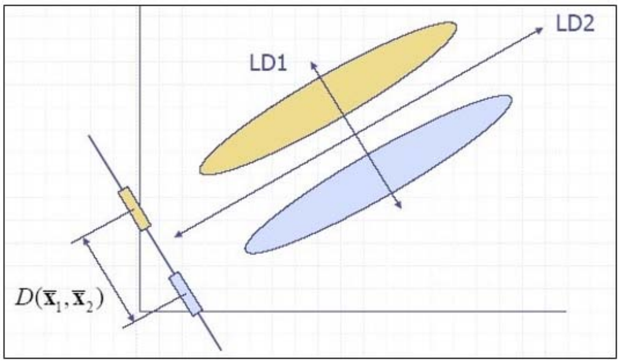


Figure 1. Graphical illustration of discriminant analysis for two categories with two discriminant functions

Discriminant analysis can be used in authorship attribution since the problem is similar to that of classifying plants into species based on their physical features. In attribution, the process is to use a set of texts of known authorship to determine the discriminant functions using non-contextual words as the features, and then classify texts of unknown authorship into the set of authors using the discriminant functions.

A variation of discriminant analysis called stepwise discriminant analysis (SDA) first determines a subset of the most distinguishing features from a comprehensive list of features and then formulates the discriminant functions (Goldstein and Dillon, 1977). The most distinguishing features are the best predictors for classifying objects into the proper categories.

In our stylometric work we have observed the utility of SDA to choose the words to use as distinguishing features in authorship attribution studies. This observation agrees with work done by other researchers (Smith and Aldridge, 2011). From a comprehensive list of non-contextual words, SDA identifies the most discriminating word first and subsequent words in descending order of discriminating ability. It stops when none of the remaining words add to the discriminating ability of the set of words. Thus, we end up with a subset of words that are the best predictors of authorship.

Another approach often used to select distinguishing features for authorship attribution is to use a list of the most frequent words listed in descending order of frequency in a set of representative texts of the candidate authors' works. Consequently, we considered this research question:

For a given set of authors and representative texts, and using Burrows' Delta as the analytical method, will the most distinguishing words (MDWs) identified by SDA give more distinguishing power in the analysis than using the most frequent words (MFWs) approach?

The corresponding null and alternative hypotheses are:

H_0: Using non-contextual MDWs selected by SDA is not more distinguishing among candidate authors than using a set of MFWs.

H_a: Using non-contextual MDWs selected by SDA is more distinguishing among candidate authors than using a set of MFWs.

Method

To answer our research question, the metric we used for a set of words' distinguishing power was the difference in Burrows' Deltas for the two authors with the smallest Deltas to that text. If the null hypothesis is true, the differences between Deltas should be about the same whether using MDWs or using MFWs. If using MDWs produces larger Delta differences than using MFWs, that evidence would support the alternative hypothesis.

We used the difference in Deltas between the nearest authors because it is an indication of statistical power. Analogous to the power of a microscope, statistical power is a statistical technique's ability to distinguish between things that are close together. The greater the distance between Deltas, the greater the power of the technique used to calculate the Deltas.

To conduct the study we used *The Federalist Papers*, commonly used for testing the usefulness of authorship attribution methods. *The Federalist Papers* are well suited to the problem as there were a total of 85 published papers written by Alexander Hamilton, James Madison and John Jay. Fifty-one were known to have been written by Hamilton, fourteen by Madison, five by Jay, and three written jointly by Hamilton and Madison. Twelve had disputed authorship, but have subsequently been studied extensively and are commonly attributed to Madison.

Because the attribution of the disputed papers is relatively non-controversial, *The Federalist Papers* provide a useful basis for comparing attribution methods. Since our research objective was not to answer the attribution question, but rather to compare methods of answering the question, using *The Federalist Paper* removed the question of correct attribution for a more direct comparison of the distinguishing ability of SDA-selected MDWs compared to MFWs.

Using only the 70 papers of known authorship as the representative texts, we applied SDA and selected the MDWs from a large list of non-contextual words, and then calculated Burrows' Delta distances for each paper to each of the three candidate authors. We compared these results to the results of using sets of MFWs ranging from 50 to 500 words in increments of 50 words.

Results

The SDA procedure select 29 words as the most distinguishing words for *The Federalist Papers*. Those 29 MDWs produced 100% correct classification of the 70 representative texts and provided greater distinguishing power than MFWs for the 12 disputed texts. As shown in figure 2, for *The Federalist Papers*, MDWs have from 1.5 to 4 times the discriminating power of MFWs.

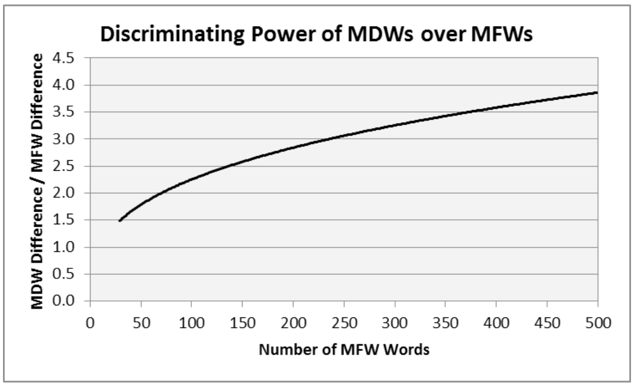


Figure 2. Comparison of the discriminating power of MDWs vs. MWFs

To understand why this occurs, examine table 1 and notice where each of the 29 MDWs appears on the MFW list.

Feature Words	MDW Rank	MFW Rank	Feature Words	MDW Rank	MFW Rank	Feature Words	MDW Rank	MFW Rank
and	1	4	the	11	1	from	21	23
upon	2	53	matter	12	364	for	22	18
whilst	3	1231	notwithstanding	13	983	any	23	41
by	4	13	also	14	241	thereby	24	3585
it	5	9	before	15	406	even	25	122
which	6	11	against	16	69	no	26	47
just	7	278	lest	17	6598	him	27	236
particularly	8	715	into	18	72	nay	28	1677
on	9	28	there	19	48	one	29	45
a	10	6	out	20	213			

Table 1. Comparative ranking of MDWs and MWFs for *The Federalist Papers*

Words in a list of MDWs often are not included in typical MFW lists. For example, note that the word, *whilst*, is the third most discriminating word selected by SDA and yet it is not even in the top 1000 MFWs. Even though Mosteller and Wallace identified *whilst* as a key indicator of authorship for the disputed papers, MFW lists of less than 1231 words would miss this highly distinguishing word. Notice further that 12 of the 29 MDWs are not even in the top 200 MFWs. So using MFWs will miss many highly distinguishing words.

Discussion

Thus, we reject the null hypothesis and assert that MDWs provide more distinguishing power between the Deltas for the two closest authors to the texts to be attributed as compared to MFWs. Our results show that MDWs can provide greater sensitivity than MFWs in discovering stylistic word-choice differences among candidate authors. The finding that it only takes 29 words selected by SDA to correctly classify all of the disputed Federalist Papers is a striking example of the power of using SDA-selected MDWs, since over 350 MFW words – more than ten times as many words – were required to achieve the same results.

Although some research has shown that variations of Delta may perform better than Burrows' original formulation (Evert et al., 2015; Hoover, 2004), we have found that using modifications of Delta does not improve the performance of MFWs relative to MDWs.

Conclusion

We conclude greater discriminating power can be achieved with a small set of MDWs chosen by SDA than with even large sets of MFWs. Using SDA-selected MDWs a researcher is more likely to make correct attributions and may be able to do it with fewer representative texts and for smaller texts. As a result, a researcher will have a greater likelihood of discovering new insights about the possible authorship of unattributed or disputed texts.

Bibliography

Argamon, S., et al. (2003). Gender, genre, and writing style in formal written texts. *Text,* **23**(3): 321-46.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing,* **17**(3): 267-87.

Evert, S., et al. (2015). Towards a better understanding of Burrows's Delta in literary authorship attribution. *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature,* Denver, CO.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* **7**(2): 179-88.

Goldstein, M. and Dillon, W. R. (1977). A stepwise discrete variable selection procedure. *Communications in Statistics – Theory and Methods,* **6**(14): 1423-36.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing,* **13**(3): 111-17.

Hoover, D. L. (2004). Delta prime?, *Literary and Linguistic Computing,* **19**(4): 477-95.

Mosteller, F. and Wallace, D. L. (2007). Inference and Disputed Authorship:

The Federalist, *The David Hume Series Philosophy and Cognitive Science Reissues*, CSLI Publications.

Smith, P. W. H. and Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. *Journal of Quantitative Linguistics,* **18**(1): 63-88.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology,* **60**(3): 538-56.

File Formats for Archiving: Stability and Persistence Issues

Peter R. Fornaro
peter.fornaro@unibas.ch
University of Basel, Switzerland

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
University of Basel, Switzerland

Introduction

A well chosen file format is an important aspect in the process of archiving digital information. A typical example are image files for archival purposes. Unfortunately it can be shown, that even a well-defined file format like TIFF can have partially proprietary, vendor-based features. Today it is inevitable to digitize analogue photographs or images because the originals are endangered by unstoppable physical decay. Because of the continuous process of scanning, digital images are an important part of our cultural heritage already and they account for a constitutive part of our contemporary multimedia output in social, scientific and economic ambits (Kuny et. al., 1998). Unfortunately any digital object must be migrated in periodic cycles, because of technological changes. Hardware migration is only one of the steps needed to ensure that a digital file can be rendered in future. The file format definition is of the same importance. If such a definition becomes obsolete, the existing file must be converted into another one that is not in danger to be outdated (Heath et. al., 2011). Therefore it is necessary to judge and optimize all relevant factors that define the sustainability of a file format definition. Even more important is the persistence of a file format if a storage solution with very long cycles of migration is used, as presented in the Monolith-Project (Fornaro et. al., 2014) or the Rosetta-Project. In those cases a lifecycle of several decades can be expected.

We discuss in this document the long-term stability of existing image file formats and derive possible new approaches. We will show in detail what weaknesses exist, that endanger the future rendering of the content. In addition an image file format definition for archival needs is proposed, based on the already existing widespread standard TIFF. The proposed approach follows the concept of the Portable Document Format (Oettler et. al.,2013), PDF and its archival derivative PDF/A. The recommended specification is called TI/A, *Tagged Image for Archives*.

Problem

If digital file formats are not well chosen, the content won't be accessible in future because it can no longer be decoded as a consequence of one or multiple technological issues (Rothenberg, 1995). A "file format" basically defines the logical structure and meaning of the bits within the bit stream and thus it is essential for correct interpretation and proper rendering of the coded data. Unfortunately a file format or parts of its logical structure and definition can become obsolete, like hardware does. As a result the information renders useless, even-though the bit stream is still properly readable. To prevent such obsolescence the file format must be migration. In most cases this is more complex than creating a simple duplicate of a bit stream; the file has to be restructured. In addition every migration can reduce image quality or introduce artefacts. Therefore it is necessary to use a file format for long-term preservation of digital data that is stable, simple, well-documented and reliable. Unfortunately most image file formats do not fulfill the needed requirements. Even open standards like the Tagged Image File Format (TIFF) can have partially vendor specific, proprietary content that decreases long-term stability. TIFF is one of the most widespread formats used to represent high quality image data in archives. TIFF is a well known, established, flexible, adaptable file format for handling images and data within a single file, by including various header tags. TIFF offers some features that are rarely used and not supported by most applications. The TIF format is quite complex and parts of the original definition have become obsolete, while new, not formally standardized additions have been made. As a consequence it would be easily possible to create a TIFF file that is fully conformant to the TIFF Revision 6.0 specifications but would be virtually useless because no existing software is be able to open and render it, migration is the needed.

Migration is an expensive task. Therefor numerous approaches for archival storage solutions exist that do not or only rarely need to be migrated. Most of those solutions make use of a very stable carrier and a simple interface to access data. Monolith is such an example for an "eternal" storage. Monolith (Gubler et. al., 2006) is based on chromogenic optical film, that has a life expectancy of up to 500 years. The data is stored on Monolith as 2D-barcode, enriched with human readable metadata. For such a solution the data format is of very high importance because any format obsolescence reduces migration cycles drastically.

Approach

Since a digital archive has the goal that the file can be rendered in a indefinite but possibly far future, a simplistic approach is necessary. Therefor a TIFF suitable for long term archiving should require only a minimal set of features (tags) that are necessary to allow a correct future rendering of the data and the essential descriptive metadata. We therefore propose a subset of the full functionality of TIFF that is fully compatible with the de-facto TIFF standard itself but marks some tags as **mandatory**, some as **optional** and some as **forbidden** in order to guarantee the correct rendering in the future. In addition to the core functionalities, it is crucial to define a minimal set of metadata for archival applications, following standards like Dublin Core or METS (Loeffler et. al., 2007). In analogy to PDF/A format we propose to call this specification *TI/A* or *Tagged Image for Archives*.

In cooperation with the University of Girona in Spain and EASY INNOVA, a technology and innovation centre of Girona, we have started the process of specifying TI/A in co-operation with multiple memory institutions of Switzerland and Europe.

Of course the concept of using a subset of the functionality of TIFF can be applied to any other format common for archiving digital image data like JPEG2000/A (Buckley, 2013) or even video or motion picture like DCP/A (Fornaro et. al., 2014: Goethels, 2009).

Results

It can be shown that a smart chosen file format is very important for successful archiving [9]. With the help of numerous institutions and experts we have drafted a recommendation, based on the existing TIFF standard (http://ti-a.org). The exchange of needs, requirements, dos and don’ts will lead us to a final draft specification of an ideal archival file format for high quality image data that is well supported by an international network of experts. Following the original standard definition of TIFF allows us to define a format that is fully compatible with existing decoders. This approach makes it not necessary to have “out on the market” software modified or enhanced by any means.

Based on that preliminary work we will try to have the document standardized by the International Standard Organisation, ISO. Such a precise definition of the functionalities and their implementation in a Tagged Image File for Archives will help to increase the sustainability of the original image format drastically.

Bibliography

Kuny, T. (1998). *A digital dark ages? Challenges in the preservation of electronic information*. Int. Preserv. News, pp. 8–13.

Rothenberg, J. (1995). *Ensuring the longevity of digital documents*. Sci. Amer. **272**(1): 42–47.

Gubler, D., Rosenthaler, L. and Fornaro, P. (2006). The obsolescence of migration: Long-Term storage of digital code on stable optical media. In *Proceedings of IS&T’s Archiving Conference*. IS&T, pp. 135–39.

Loeffler, H. (2007). Photo Metadata White Paper. In Baranger, W. (Eds.),*IPTC,* http://www.iptc.org/std/photometadata/0.0/documentation/IPTC-PhotoMetadataWhitePaper2007_11.pdf (accessed 4. March 2016)

Goethels, A. (2009). *General Considerations for Choosing File Formats*, Harvard University Library, http://library.harvard.edu/sites/default/files/general_format_considerations.pdf (accessed 4. March 2016).

Heath, T. and Bizer, Ch. (2011). *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool.

Oettler, A. (2013). *PDF/A in a Nutshell 2.0,* – presentation from the First International PDF/A Conference in Amsterdam, Association for Digital Document Standards e.V., Berlin.

Buckley R. (2013). *Using Lossy JPEG2000 Compression For Archival Master Files*, Library of Congress Office of Strategic Initiatives, Version 1.1, http://www.digitizationguidelines.gov/still-image/documents/JP2LossyCompression.pdf (accessed 4. March 2016).

Fornaro, P. and Gubler D. (2014). *DCP/A: Discussion of an Archival Digital Cinema Package for A V-Media*, IS&T Archiving Conference Proceedings, Berlin.

Fornaro P., Wassmer A., Rosenthaler L. and Gschwind R. (2014). *Monolith: Materialised Bits, the Digital Rosetta Film*, DH2014 Conference, Lausanne.

Approaches to Thematic Classification for Latin Epic

Christopher W. Forstall
forstall@buffalo.edu
Université de Genève, Switzerland

Lavinia Galli Milic
lavinia.gallimilic@unige.ch
Université de Genève, Switzerland

Nelis Damien
damien.nelis@unige.ch
Université de Genève, Switzerland

Background and Motivation

It has long been understood that Greek and Roman epic poems partake of a shared repertory of stock themes and typical scenes: the catalogue of heroes, the warrior arming for battle, the tempest, etc. These themes originally evolved under circumstances peculiar to oral-formulaic composition in archaic Greece, where they served the exigencies of real-time composition in performance by organizing poetic material into mnemonic chunks, each with a predictable internal structure (Rubin, 1995; Minchin, 2001). For later, literate Greeks, and still later for the Romans, who continued to develop the epic tradition, the use of these themes was no longer an aide to memory, but a complex intertextual gesture that formed one of the defining features of the genre (De Jong, 2014; Nünlist, 2009).

The study of elemental, thematic building-blocks in epic and related genres has a long history, through the twentieth century and even earlier, including work by Claude Levi-Strauss, Milman Parry, and Vladimir Propp. While catalogues and typologies of epic themes exist (*e.g.*, Edwards, 1992), at the same time significant disagreement remains over their definition and delineation. Here we take first steps towards automated detection of theme in epic, defining features that target scene-sized samples of text under in bag-of-words model, and applying both unsupervised and supervised classification methods.

Our research is on intertextuality in Latin epics of the Flavian period, and in particular the ways in which an intertextual relationship at the thematic level can support or undermine specific verbal allusions at the sentence or phrase level. The ability to automatically detect text reuse in Latin epic at the scale of individual verse lines is provided by tools such as Tesserae (http://tesserae.caset.buffalo.edu), Musisque Deoque's co-occurrence search (http://www.mqdq.it/mqdq/cooccorrenze.jsp), and eTRAP's forthcoming TRACER framework (http://etrap.gcdh.de), so that it is now conceivable to generate an exhaustive list of all such correspondences between any two texts. Yet the same efforts have shown that simple text-reuse search alone cannot capture all of the allusions noted by professional commentaries, and that sensitivity to scene-level thematic parallelism would improve both recall and precision over a model based on single verse lines or sentences (Coffee et al., 2012).

Tesserae has in fact produced a prototype thematic search using topic modelling to create scene-level features (Scheirer et al., forthcoming). At the same time, work by the Memorata Poetis project affiliated with Musisque Deoque is carrying out systematic manual tagging of themes in Latin vernacular poetry (Ciotti et al., 2015). Both approaches—the supervised and the unsupervised identification of themes—show promise, although neither has been fully integrated into its respective parent project's verse-level, text-reuse search tool. In the ongoing work presented here, we attempt to combine elements of each approach, comparing unsupervised, bag-of-words classifiers at the scene-level with manual tagging for specific, selected themes. Our ultimate goal is to combine similarity scores for thematic parallelism with existing phrase-level search tools' scores for text reuse in order to improve their accuracy.

Method

We consider a corpus including the three more or less complete epics of the Flavian period—Valerius Flaccus' *Argonautica*, Statius' *Thebaid*, and Silius Italicus' *Punica*—as well as three earlier poems to which our works of interest respond—Lucan's *Civil War*, Ovid's *Metamorphoses*, and Vergil's *Aeneid*. These works are initially subdivided into samples of 50 consecutive verse lines. After lemmatization, tf-idf weighted feature vectors are calculated for each sample, dropping terms common to 50% or more of the samples, and the resulting feature set is reduced using principal components analysis to the most significant 500 components.

The samples still show a strong tendency to cluster by author despite the removal of very frequent words. Since our goal is to examine parallel variation across authors, we attempt to remove a characteristic 'authorship signal' from each author's *œuvre* before classification. A mean vector representing the author is thus subtracted from every one of his samples.

In the unsupervised approach, we then perform k-means clustering on the corpus as a whole. Because we do not have any *a priori* set of themes to look for, we test multiple values of *k* in an attempt to identify the most stable number of clusters. For each possible value of *k*, we perform multiple repetitions and compare pairwise agreement among the resulting classifications using the adjusted Rand index. We also attempt to identify the most stable passages through comparison of error across repeated classifications using different offsets for our 50-line sampling window. For those passages most consistently classified across repetitions, we return to the text to investigate, through close reading, whether the passages indeed demonstrate meaningful similarity.

In the supervised approach, on the other hand, we begin by identifying scenes of interest using a set of *a priori* thematic categories—for example, battle scenes and storms at sea. We then attempt to train a classifier that can distinguish these types. Earlier work on this project has shown that certain thematic contexts can be separated at a coarse scale (e.g, love versus war, generally) using principal components analysis alone; however here we will attempt to achieve finer precision using support vector machines, a popular approach in stylometric tasks and one that has previously been applied to intertextuality in Latin in particular (Forstall et al., 2011).

Results

This work is ongoing, and one of our goals in presenting at DH 2016 is to elicit feedback that will help to shape the design of future experiments. At the same time, early results suggest that the supervised and unsupervised approaches may indeed meet in the middle. For example, many of the most stable passages in the unsupervised classification belong to a 'nautical' theme which includes as a subset the storm theme identified by human tagging. Across 100 different classifications of the six epics, using differently-offset sampling windows, the passages with the lowest rate of disagreement included characteristic descriptions of tempests in *Aeneid* 1 and *Argonautica* 1 (Figs. 1–2). The same class, and a high degree of stability, also marked nautical chase sequences in *Aeneid* 5 and *Argonautica* 8 as well as the recounting of the Argonauts' story in *Thebaid* 5.

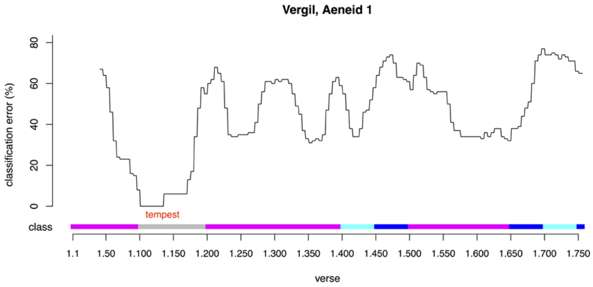


Figure 1. Vergil's *Aeneid*, book 1, after k-means classification with 8 classes. The line shows disagreement between repeated re-classifications. Shaded bands at bottom show classes in one randomly-selected trial.

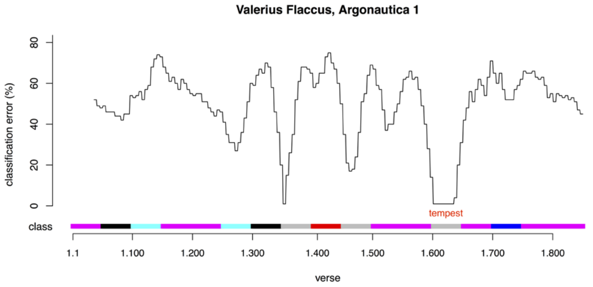


Figure 2. Valerius Flaccus, *Argonautica*, book 1, after k-means classification with 8 classes. Line shows disagreement among repeated re-classifications; shaded bands show the same classes as in Fig. 1.

Implications

The concrete goal of this project is to provide a thematic feature set for automated intertext-detection, compatible with the phrase-based results of existing tools, so that, for example, otherwise slight verbal correspondences can be promoted where they occur within thematically parallel passages and thus are likely more interesting to readers. More generally, we would like to understand how stock themes function in epic intertextuality, and whether they can indeed be sufficiently modeled using a bag-of-words approach. To the degree that they can, we can say that thematic features are in fact made of words and exist on a continuum with the word-level bigrams matched by Tesserae, for example, instead of representing 'deeper', semiotic structures distinct from the word forms that appear on the 'surface' of the text (as in, e.g., Levi-Strauss, 1955)

Bibliography

Ciotti, F., Mordenti, R., and Silvi, D. (2015). Thematic Annotation of Literary Text: The Case for Ontology. Paper presented at *Humanités Numeriques et Antiquité*, Grenoble, France.

Coffee, N., et al. (2012). Intertextuality in the Digital Age. *Transactions of the American Philological Association*, **142**(2): 383–422.

De Jong, I. J. F. (2014). *Narratology and Classics: A Practical Guide*. New York: Oxford University Press.

Edwards, M. W. (1992). Homer and Oral Tradition: The Type-Scene. *Oral Tradition*, **7** (2): 284–330.

Forstall, C. W., Jacobson, S. L., and Scheirer, W. J. (2011). Evidence of intertextuality: investigating Paul the Deacon's *Angustae Vitae*. *Literary and Linguistic Computing*, **26** (3): 285–96.

Levi-Strauss, C. (1955). The Structural Study of Myth. *The Journal of American Folklore*, **68** (270): 428–44.

Minchin, E. (2001). *Homer and the Resources of Memory*. New York: Oxford University Press.

Nünlist, R. (2009). *The Ancient Critic at Work: Terms and Concepts of Literary Criticism in Greek Scholia*. Cambridge: Cambridge University Press.

Rubin, D. C. (1995). *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. New York: Oxford University Press.

Scheirer, W. J., Forstall, C. W., and Coffee, N. (forthcoming). The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning. *Digital Scholarship in the Humanities*. First published online October 20, 2014: 10.1093/llc/fqu058.

Toccata : Text-Oriented Computational Classifier Applicable To Authorship

Richard Sandes Forsyth
forsyth_rich@hotmail.com
independent, United Kingdom

Introduction

Many text-classification techniques have been proposed and used for authorship attribution (Holmes, 1994; Grieve, 2007; Juola, 2008; Koppel et al., 2011), genre categorization

(Biber, 1988; Argamon et al., 2003), stylochronometry (Forsyth, 1999) and other tasks within computational stylistics. However, until quite recently, it has been extremely difficult to assess novel and existing techniques on comparable benchmark problems within a common framework using statistically robust methods.

Toccata is a resource for computational stylometry which aims to address that lack, freely available at http://www.richardsandesforsyth.net/software.html under the GNU public licence.

The main program is a test harness in which a variety of text-classification algorithms can be evaluated on unproblematic cases and, if required, applied to disputed cases. The package supplies four pre-existing classification methods as modules (including Delta (Burrows, 2002), widely regarded as a standard in this area) as well as five sample corpora (including the famous *Federalist Papers*) so that users who don't wish to write Python code can use it simply as an off-the-shelf classifier and those who do can familiarize themselves with the system before implementing their own algorithms.

Noteworthy features of the system include:

1. sample corpora provided for familiarization;
2. test phase using random subsampling to give robust error-rate estimation;
3. ability to plug in new techniques or to employ existing standards;
4. option of post-hoc phase applying trained model(s) to unseen holdout data;
5. empirically grounded computation of post-hoc confidence weights to deal with 'open' problems where the unseen cases may not belong to any of the training-set categories;
6. accompanying export file readable by R or similar statistical packages for optional further processing.

Sketch of the System's Operation

Toccata performs three main functions, in sequence:

(a) testmode: leave-n-out random resampling test of the classifier on the training corpus to provide statistics by which the classifier can be evaluated;

(b) holdout: application of the classifier to an unseen holdout sample of texts, if given;

(c) posthoc: re-application to the holdout sample of texts (if given) using the results from phase (a) to estimate empirical probabilities.

Steps (b) and (c) are optional.

Sample corpora

Toccata is a document-oriented system. Thus a training corpus consists of a number of text files, in UTF8 encoding, without markup such as HTML tags. Each file is treated as an individual document, belonging to a particular category. Example corpora are supplied to enable users to start using the system, prior to collecting or reformatting their own corpora.

ajps: ninety poems by 2 eminent 19th-century Hungarian poets, Arany József and Petőfi Sándor. Arany was godfather to Petőfi's child, so we might expect their writing styles to be relatively similar.

cics: Latin texts relevant to the authorship of the *Consolatio* which Cicero wrote in 45 BC. This was thought to have been lost until in 1583 AD when Sigonio claimed to have rediscovered it. Background information can be found in Forsyth et al. (1999).

feds: writings by Alexander Hamilton and James Madison, as well as some contemporaries of theirs. This corpus is related to another notable authorship dispute, concerning the *Federalist Papers*, which were published in New York in 1788. See Holmes and Forsyth (1995).

mags: 144 texts from 2 different learned journals, namely *Literary and Linguistic Computing* and *Machine Learning*. Each text is an excerpt consisting of the Abstract plus initial paragraph of an article in one of those journals, written during the period 1987-1995.

sonnets: 196 English sonnets, 14 each by 14 different authors, with an additional holdout sample of 24 texts, half of which are by authors absent from the main sample.

Validation by Random Subsampling

A major objective of the system is to assess the effectiveness of text-classification methods by a form of cross validation. For this purpose the training corpus of undisputed texts is repeatedly divided into two portions, one used to form a classification model and the other used to test the accuracy of this model. After this cycle a number of quality statistics are computed and printed, along with a confusion matrix. This helps to establish a relatively honest estimate of the likely future error rate of the classifier. After subsampling, the program will construct a model on the full training set. This may then be applied to a genuine holdout sample, if provided.

Classifier Modules

A classifier module is expected to develop trained models of each text category and deliver matching scores of a text to each model, with more positive scores indicating stronger matching. The category with the highest match-score relative to the average of all scores for the text, is the assigned class. Four library modules are supplied "off the shelf".

Module **docalib_deltoid.py** is an implementation of Burrows's delta (Burrows, 2002) which has become a standard technique in authorship attribution studies. Module **docalib_keytoks.py** works by first finding the 1024 most common word tokens in the corpus, then keeping from these the most distinctive. For classification, relative word frequencies in the text being classified are correlated with relative frequencies in each class. Module **docalib_maws.py** is a version of what Mosteller and Wallace in their classic work (1964/1984) on the *Federalist Papers* call their "robust Bayesian analysis", as implemented by Forsyth (1995). Module **docalib_topvocs.py** implements another classifier inspired by the approach of Burrows (1992), which uses the most frequent tokens in the training corpus as features.

The Holdout and Posthoc Phases

The subsampling test phase (above) is primarily concerned with assessing the quality of a classification method. The holdout and posthoc phases are when that method is applied in earnest.

If a holdout sample is given, the model developed on the training set is applied to that sample. The holdout texts may belong to categories that were not present in the training set, so each decision is categorized as correct (+), incorrect (-) or undetermined (?) and the success rate statistics computed accordingly.

This is illustrated in Table 1, below, from an application of the MAWS (Mosteller and Wallace) method to a collection of sonnets. Here the training set consists of 196 short English poems -- 14 sonnets by 14 different authors. This is a challenging problem firstly because the median length of each text in the training corpus is 116 words, secondly because 14 is a relatively large number of candidates.

Table 1 shows the ranking produced on a holdout sample of 24 texts, absent from the training set. Note that 12 of these 24 items are 'distractors', i.e. texts by authors not present in the training set. The program assigns these a question mark (?) in assessing its own decision.

The listing ranks the program's decisions from most to least credible. The upper third include 6 correct assignments, 1 clear mistake and a distractor. The middle third contains 1 correct classification, 3 mistakes and 4 distractors. The last third contains no correct answers, 1 mistake and 7 distractors. (Incidentally, the distractor poem by the Earl of Oxford, ranked twentieth, is more congruent with Wordsworth than any other author, including Shakespeare, and not confidently assigned to any of the training categories.)

This output addresses the very real problem of documents from outside the known training categories. The listing is ordered by a quantity labelled 'credit'. This is the geometric mean of the last two numbers in each line, labelled 'confidence' and 'congruity'. Confidence is derived from the preceding subsampling phase. It is computed from the differential matching score of the text under consideration as W / (W+L), where W is the number of correct answers which received a lower differential score during the subsampling phase and L is the number of wrong answers with a higher score. Congruity is simply the proportion of matching scores of the chosen category that were lower, in the subsampling phase, than the score for the case in question. It is an empirically based index of compatibility between the assigned category of the text and the training examples of that category.

In all kinds of classification, the problem of never-before-seen categories can loom large. (See, for instance, Eder, 2013.) Like most trainable classifiers, Toccata always picks the most likely category from those it has encountered in training, but the most likely may not be very likely. The confidence and congruity scores give useful information in this regard. For example, if we only consider the classifications which obtain a score of at least 0.5 on both confidence and congruity, we find 6 correct decisions, 1 incorrect and 1 distractor. Treating the distractor (assigning a sonnet by Dylan Thomas to Edna Millay) as incorrect still represents a 75% success rate in an "open" authorship problem on texts only slightly more than a hundred word tokens in length, where the training sample for each known category consists of approximately 1600 words, with a chance expectation of 7% success. In other words, three crucial parameters -- training corpus size, text length and number of categories -- are all well "outside the envelope" of most previously reported authorship studies.

Rank	credit	Filename	pred:true	conf.	congruity
1	0.9163	ChrRoss_WinterSecret.t	ChrRoss + ChrRoss	0.9530	0.8810
2	0.8768	WilShak_6.txt	WilShak + WilShak	0.9425	0.8158
3	0.8142	DylThom_Altar09.txt	EdnMill ? DylThom	0.8838	0.7500
4	0.7664	MicDray_Idea000.txt	MicDray + MicDray	0.6378	0.9211
5	0.7595	WilShak_137.txt	WilShak + WilShak	0.8118	0.7105
6	0.6950	JohDonn_Nativity.txt	JohDonn + JohDonn	0.6720	0.7188
7	0.6247	MicDray_Idea048.txt	JohDonn – MicDray	0.5430	0.7188
8	0.5356	WilShak_109.txt	WilShak + WilShak	0.5737	0.5000
9	0.5225	DylThom_Altar05.txt	RupBroo ? DylThom	0.4150	0.6579
10	0.4684	TomWyat_THEY_FLEE	EdmSpen ? ThoWyat	0.4596	0.4773
11	0.4226	PerShel_Ozymandias.txt	EliBrow ? PerShel	0.2217	0.8056
12	0.4027	EliBrow_SP23.txt	DanRoss – EliBrow	0.2237	0.7250
13	0.3061	WilShak_RomeoJuliet.tx	WilShak + WilShak	0.2094	0.4474
14	0.2739	PhiSidn_astel108.txt	EliBrow – PhiSidn	0.1080	0.6944

15	0.2625	DylThom_Altaro6.txt	EliBrow ? DylThom	0.0992	0.6944
16	0.2283	JohDonn_Temple.txt	EdnMill – JohDonn	0.1179	0.4423
17	0.2014	Lincoln-1863Gettysburg.	SamDani ? AbeLinc	0.0649	0.6250
18	0.1894	RicFors_LaBocca.txt	RupBroo ? RicFors	0.0649	0.5526
19	0.1352	HelFors_1958.txt	EliBrow ? HelFors	0.0263	0.6944
20	0.1089	oxford_13.txt	WilWord ? Oxford	0.0265	0.4474
21	0.0977	RicFors_Underworld.txt	EdnMill ? RicFors	0.0261	0.3654
22	0.0755	HelFors_1982.txt	DanRoss ? HelFors	0.0109	0.5250
23	0.0690	DylThom_Altaro3.txt	RupBroo ? DylThom	0.0106	0.4474
24	0.0411	PhiSidn_astelo30.txt	EdmSpen – PhiSidn	0.0106	0.1591

Table 1. Posthoc ranking of 24 decisions on unseen texts, including 12 'distractors'

Bibliography

Argamon, S., et al. (2003). Gender, genre, and writing style in formal written texts. *Text*, **23**(3): 321-46.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Burrows, J.F. (1992). Not unless you ask nicely: the interpretive nexus between analysis and information. *Literary and Linguistic Computing*, **7**(2): 91-109.

Burrows, J.F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267-87.

Eder, M. (2013). Bootstrapping Delta: a safety net in open-set authorship attribution. *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska-Lincoln, pp. 169-72.

Forsyth, R.S. (1995). *Stylistic Structures: a Computational Approach to Text Classification*. Unpublished Doctoral Thesis, Faculty of Science, University of Nottingham. http://www.richardsandesforsyth.net/doctoral.html

Forsyth, R.S. (1999). Stylochronometry with substrings, or: a poet young and old. *Literary and Linguistic Computing*, **14**(4): 467-77.

Forsyth, R.S., Holmes, D.I. and Tse, E.K. (1999). Cicero, Sigonio, and Burrows: investigating the authenticity of the 'Consolatio'. *Literary and Linguistic Computing*, **14**(3): 1-26.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, **22**(3): 251-70.

Holmes, D. (1994). Authorship attribution. *Computers and the Humanities*, **28**: 1-20.

Holmes, D.I. and Forsyth, R.S. (1995). The 'Federalist' revisited: new directions in authorship attribution. *Literary and Linguistic Computing*, **10**(2): 111-27.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233-334.

Koppel, M., Schler, J. and Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, **45**, pp. 83-94. DOI 10.1007/s10579-009-9111-2.

Mosteller, F. and Wallace, D.L. (1984). *Applied Bayesian and Classical Inference: the Case of the Federalist Papers*. New York: Springer. [First edition, 1964.]

Repairing William Playfair: Digital Fabrication, Design Theory, and the Long History of Data Visualization

Caroline Rebecca Foster
cfoster2@gatech.edu
Georgia Institute of Technology, United States of America

Erica Pramer
erica.pramer@gatech.edu
Georgia Institute of Technology, United States of America

Lauren Klein
lauren.klein@lmc.gatech.edu
Georgia Institute of Technology, United States of America

Introduction

In her plenary address at the 2014 Digital Humanities conference, Bethany Nowviskie urged the field to consider how “broken world thinking,” an approach equal parts ethical, ontological, and methodological, might enrich digital humanities practice (2015, n.p.). Nowviskie borrows the phrase from information theorist Steven Jackson, who argues for a reparative rather than productivist approach to the study of media and technology, and more specifically, for an increased emphasis on the “moments of breakdown” that allow us to “see and engage our technologies in new and sometimes surprising ways” (2013, 230). In this paper, we take up this shared call and extend it, elaborating an approach to broken-world thinking that is simultaneously informed by examples of historical fabrication in the digital humanities (e.g. Elliott et al. 2012, Sayers 2015) and theories of breakdown and repair from the field of design (e.g. Jackson, 2013; Gabrys, 2011). We take the time-series charts of William Playfair, the eighteenth-century data visualization pioneer, and recreate them using D3.js, a data visualization library commonly employed in digital humanities work (Bostock, 2011). In doing so, we gain valuable purchase on the historical concepts that contributed to the creation of Playfair's charts, many of which-- such as data-- still hold sway today. But by remaining equally attentive to the "moments of breakdown" between the original artifact and our contemporary recreations, we are also able to open new perspectives on the "affordances" of our own visualization tools (Murray, 2011). Our digital "fossils," as we term them, following the work of Jennifer Gabrys, suggest a generative new point of intersection between the fields of digital humanities and design (2011).

Project overview

William Playfair is widely considered the "inventor" of modern data visualization (Tufte, 32). The graphical forms that he first developed, including the bar chart and the pie chart, remain among the building blocks of visualization today (Wilkinson, 2005), and the charts he created are still employed as examples of the crystallizing power of data visualization (Klein, forthcoming). And yet, the techniques he employed, such as copperplate engraving, have long been supplanted by newer technologies. This project seeks to understand how Playfair's techniques affected the images he created, and how our techniques, in turn, affect the images and interactions we create today.

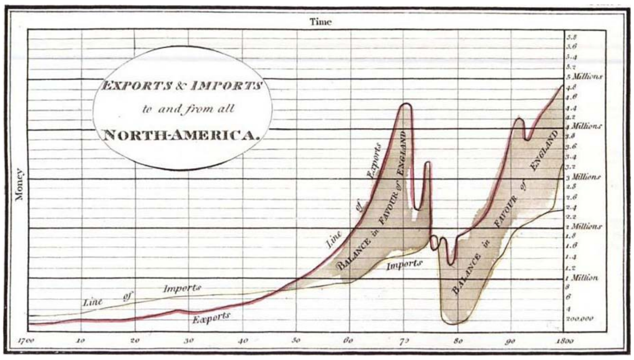


Figure 1: William Playfair, "Exports and Imports to and from all North-America," from *The Commercial and Political Atlas*, 3rd ed. London, 1801

To recreate Playfair's chart, we selected D3, the javascript visualization library employed in contexts ranging from data journalism to scientific research to the digital humanities (e.g. Meeks n.d., Schmidt n.d). In comparison to off-the-shelf software such as Microsoft Excel or Tableau, D3 provides additional control over the structure and style of the data, an advantage when attempting to achieve fidelity to an original image. In addition, D3 is open source; this allowed us to consider additional aspects of the library's design. Finally, D3 was developed in an academic context; its own design choices therefore support a conceptual as well as technical analysis.

We took two approaches to recreating Playfair's chart: the first by adhering to the original as closely as possible, including the use of the original data; and the second by adapting Playfair's design for use with contemporary trade data, taking advantage of D3's emphasis on data transformation. (We employed the US Census Bureau's data on foreign trade). In the sections that follow, we describe these approaches in more detail, with particular attention to "moments of breakdown" and the new perspectives that they granted.

First approach: remediating playfair's original chart as a digital fossil

Jennifer Gabrys, a design theorist who, like Jackson, views instances of breakdown and failure in a generative light, suggests that we view cast-off objects as "fossil forms" rather than waste (2011, 7). These digital fossils provide "evidence of more complex and contingent material events," as well as "traces of the economic, cultural, and political contexts in which they circulate." By recreating Playfair's chart in D3, we also remediate its "fossil form," granting us access to the various contexts in which the chart circulates, both historical and contemporary.

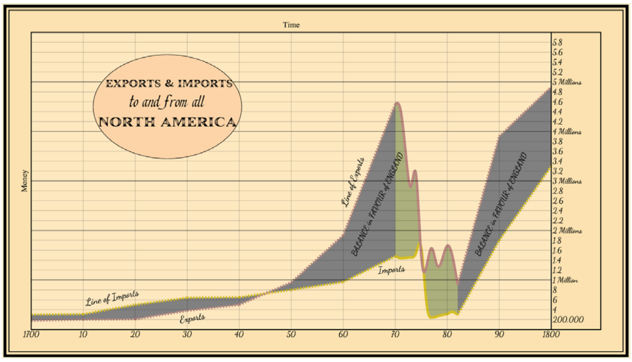


Figure 2: True-to-form recreation of William Playfair's original chart. The gray area emphasizes the uncertainty of the data, while the green area matches the original. Implementation and image by C. Foster

Our interest in creating our digital fossil was to induce the moments of breakdown that might alert us to the contextual differences between past and present; the nature and status of statistical data was one such difference. When consulting the third edition of Playfair's *Statistical Atlas* as a reference, we found no actual data accompanying his charts. D3 assumes that the developer will begin with data, so without it, we could not begin. To compensate, we turned to a data table from a previous edition, but it contained data for only a portion of the date range, from 1770 to 1782. We began by recreating that section of time, but to create the entire chart, we estimated the additional data points. The resultant chart resembled the original, but was premised upon two different data sources, with different degrees of accuracy.

This instance of breakdown and repair illustrates how D3 assumes that a dataset will be presented in a certain format, and that the data will be well-defined, clean, and accurate. The context of D3 is revealed as representative of a culture fixated on data-driven solutions. Rather than present our numbers, actual and interpreted, as the same, we used a technique developed by Kevin Schaul (2013) to create dashed lines for the interpreted numbers. The code he developed, what some might view as a "hack," might be understood as a "repair" of a breakdown within D3, one that enables the visual presentation of defined and undefined data together. By contrast, Playfair's original chart shows us that precise data were not a necessary component of its initial success. Playfair drew his charts' data lines freehand. In fact, there is little evidence that Playfair plotted any actual data points before engraving the lines (Klein, forthcoming).

Second approach: creating an interactive chart in the style of playfair

Our second recreation, an interactive version of Playfair's chart supplemented by modern trade data, revealed additional contexts and biases encoded in D3's design. D3 was designed to facilitate the creation of interactive visualizations (Bostock, 2011). Its built-in functions worked smoothly once we traded out the original dataset for a more consistently formatted, if substantially larger, contemporary one.

It was when we attempted to recreate Playfair's customized labels that we encountered a significant moment of breakdown. All Playfair did to create his labels was to pick an appropriate spot and engrave them. While less extensible than any computational method, Playfair's technique allowed for more flexibility in the visualization's layout. Since we were dynamically generating the charts, we weren't able to use the human eye. Instead, we had to determine a set of rules for where to place text, and then encode them in D3. To ensure legibility, we had to verify three things: 1) that the label was not placed on a part of the chart where the import and export lines were too close; 2) that the label did not intersect with a line; and 3) that the text was placed along a part of the graph that had a consistent slope. As it turned out, determining the points of intersection was a non-trivial task. Even though the ability to illustrate the intersections between lines-- or more generally, the relations among different slices of a particular dataset-- would seem to be a basic requirement of any visualization platform, D3 was constrained by the affordances of its underlying technologies. Playfair thought hard about how to facilitate a "comparative perspective" through the design of his charts, but employing contemporary tools that are constrained for various reasons can affect the range of knowledge that is produced (1801, x).

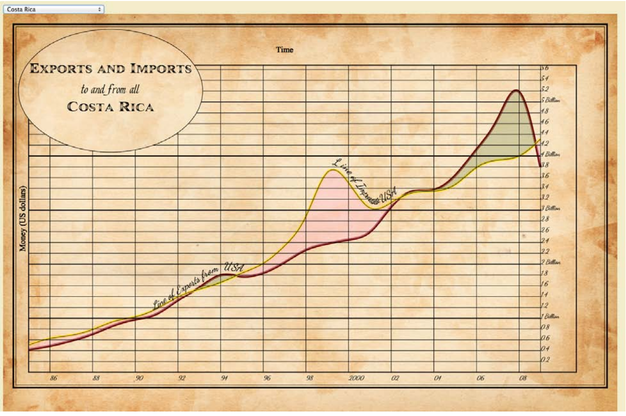


Figure 3: Interactive version of Playfair's time-series charts. The user selects the country to display through a drop-down menu. Implementation and image by E. Pramer

Conclusions

This project illustrates some of the insights that emerge from broken-world thinking as applied to digital humanities tools. Through the process of recreating Playfair's charts, we introduced moments of breakdown and prompted our repairs. We became alerted to the changed relation between data and image, and to how the hidden affordances of both software and platform affect the forms of knowledge that D3 can produce. Copperplate engraving allowed greater flexibility and less reliance on the dataset. By contrast, D3 imposes limits on design and is heavily reliant on a clean dataset. This project shows how an uninterrogated reliance on popular tools can limit the creative expression of humanistic data. We have since extended this study by recreating the visualizations Elizabeth Peabody (1804-1894). Her visualization techniques are far more difficult to recreate using standard tools, underscoring how historical fabrication allows us not only to better understand the past, but also to illuminate the present.

Bibliography

Bostock, M.,Ogievetsky, V., and Heer, J. (2011). D3: Data-Driven Documents, *IEEE Transactions of Visualization and Computer Graphics,* **17**(12): 2301-09.

Elliott, D., MacDougall, R. and Turkel, W. J. (2012). New Old Things: Fabrication, physical computing, and experiment in historical practice, *Canadian Journal of Communication,* **37**(1). Web.

Gabrys, J.. (2011). *Digital Rubbish: A natural history of electronics,* University of Michigan Press, Ann Arbor, MI.

Jackson, S. J. (2013). Rethinking Repair, in T. Gillespie, P. Boczkowski, and K. Foot, *Media Technologies: Essays on Communication, Materiality, and Society,* MIT Press, Cambridge, MA, pp. 221-39.

Klein, L. F. (forthcoming). 'Data, Image, and D3: Recreating William Playfair,' in S. Jackson et al., eds, *DigitalSTS: A Handbook and Field Guide*. University of Chicago Press, Chicago, IL.

Meeks, E. (n.d). *Digital Humanities Specialist*, Stanford University Libraries, accessed at: https://dhs.stanford.edu/category/d3/ (28 October 2015).

Murray, J. (2011). *Inventing the Medium: Principles of Interaction Design as a Cultural Practice*, MIT Press, Cambridge, MA.

Nowviskie, B. (forthcoming). Digital Humanities In The Anthropocene. *Digital Scholarship in the Humanities*.

Playfair, W. (1801). *The Commercial and Political Atlas*, 3rd ed., London.

Rosner, D. K. and Ames, M. (2014). Designing for Repair? Infrastructures and Materialities of Breakdown, *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* - CSCW '14.

Sayers, J. (2015). Why Fabricate?, *Scholarly and Research Communication*, **6**(3), n.p.

Schaul, K. (2013). *Tutorial: Undefined data in d3 charts,* accessed at: http://kevin.schaul.io/2013/07/06/undefined-data-in-d3-charts/ (28 October 2015).

Schmidt, B. (n.d.). *Maps and Visualization Gallery*, accessed at: http://benschmidt.org/maps-visualizations-gallery/ (28 October 2015).

Tufte, E. (2001). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.

Wilkinson, L. (2005). *The Grammar of Graphics,* Springer-Verlag, New York, NY.

Anonymity and Online Discussion: A New Framework for Analysis

Rolf Fredheim
ref38@cam.ac.uk
CRASSH, University of Cambridge, United Kingdom

Alfred Moore
am2214@cam.ac.uk
CRASSH, University of Cambridge, United Kingdom

We present a way of disaggregating the concept of anonymity and linking it to particular democratic goods. We demonstrate its value in the particular context of studying online commenting by contrasting three distinctive commenting regimes used by the Huffington Post (HuffPo) in the period January 2013 - March 2015. But we think this may have wider relevance for scholars looking at online communication, as modes of identity disclosure are an unavoidable yet consequential design feature of online communication platforms. In this paper we disaggregate three aspects of anonymity: traceability, which we argue relates to inclusion; durability, which influences deliberative quality through varying levels of communicative accountability; and connectedness, which ties users to real-world relations.

Decisions about online architectures have a crucial influence on the quality of communication, but are often adopted by default or with regard to other factors. In the realm of online commenting, we observe a trend of commenting platforms being outsourced to the Facebook commenting API. These changes are often framed and justified in terms of a shift from an anonymous but easily abused environment, to one in which users operate under real-name identities. When we think in terms of a choice between anonymous and real-name architectures, it seems that there is a simple trade off between the goods and dangers associated with anonymity, such as trolling on the one hand and the freedom to express one's views without fear of recrimination on the other, and the goods and dangers of real-name communication, which ties users more closely to discursive norms of civility but which also risks reproducing offline power relations.

We propose to analyse the concept of anonymity in three dimensions. We argue that cross-platform connectedness is associated with increasingly personalised communication as well as less communicative engagement between individual users; durability (continuity) of identity is associated with greater levels of civility, reduced trolling, and higher quality of deliberation (more reason-giving, etc); traceability is associated with exit - that is, people worried about traceability leads people to make a binary decision, to opt out, and the increasingly pervasive traceability promotes exit. Thinking of anonymity in these terms helps resolve some apparent trade-offs: it may be possible to capture the benefits of communicative accountability without the drawbacks of either the abusive space of easy anonymity or of the exposure to offline power relationships associated with real-name spaces.

Disaggregating anonymity

Anonymity exists on one end of a spectrum of degrees of disclosure of identity. In the context of commenting, anonymity means contributions to a discussion are not visibly linked to any particular commenter. Pseudonyms - names chosen by commenters - allow users to keep their real identity private if they wish, yet allows them to maintain a persistent alternate identity in the context of the forum in question. You might have an identity as a commenter on the *Guardian* that you keep separate from your professional networks and from your social networks. Real name comments, obviously enough, appear under your real name, not a pseudonym.

While this is a useful starting point, many scholars analysing the different degrees of disclosure on online platforms have sought more fine-grained distinctions

(Marx, 1999; Ruesch and Märker, 2012). We favour splitting online anonymity into three dimensions: traceability, durability, and connectedness.

Traceability refers to the extent to which your contributions can be traced to your real identity. Traceability is distinct from disclosure of identity to fellow commenters. You can make comments under a pseudonym and yet it may be possible (with some effort) for advertisers or the National Security Agency (NSA) to trace your real identity. Nissenbaum, for instance, argues that anonymity online, in the sense of 'conducting one's affairs, communicating, or engaging in transactions anonymously in the electronic sphere... without one's name being known', is undermined by technologies that have made it possible to track and / or piece together ('infer identity from non-identifying signs and information') the real identities of citizens online even when they are withholding their names or using pseudonyms (Nissenbaum, 1999). Because traceability is strictly distinct from the question of anonymity or pseudonymity *with respect to other commenters*, it does not help us grasp the relation between disclosure of identity and discussion quality.

Durability refers to the ease or difficulty with which online identities can be acquired and changed. Where new pseudonyms are easy to create, online identities are disposable; if you acquire a reputation for abusive or untrustworthy behaviour you can just create a new pseudonym and start again. As Resnick and Friedman put it, cheap pseudonyms create 'opportunities to misbehave without paying reputational consequences' (Resnick and Friedman, 2001). Users are more likely to stick with a particular name, exposing them to the reputational consequences of their behaviour. The durability dimension is particularly important for the possibility of holding commenters accountable for claims they make, enabling challenges in terms of consistency, and exposing uncivil or abusive commenters to sanctions.

The third dimension has to do with **connectedness** or bridging across different platforms and contexts. The most visible example of cross-platform connectedness is sites enabling users to login or register through a third party - typically Google or Facebook - rather than filling in a site-specific form. Connectedness also involves reputation, but is a global rather than local reputation. The durability or stickiness of an online identity is a necessary condition for building a local reputation, but it need not be connected to the wider reputation, a cross-referenced, cross-platform (including real life) reputation, of the sort you would want if you were renting out your apartment to someone you didn't know (as in Airbnb).

The Experiment

The two changes introduced by HuffPo provided us with a natural experiment. We collected more than 50 million comments on more than 50,000 articles featured on the HuffPo front page in the period January 2013 - March 2015.

The first of these phases allowed anonymous commenting. At this time, the platform experienced aggressive 'trolling' and the use of multiple accounts. In December 2013 HuffPo moved to regulate its forum by requiring users to authenticate their accounts through Facebook. On the face of it, little changed in this pseudonymous environment: user names and avatars remained, but behind the scenes Facebook's database helped weed out fake accounts. Users did not have to create a new online identity, give up their old pseudonyms, or change the appearance of their HuffPo commenter profile. We read this first change as primarily about durability of identity.

In June 2014 HuffPo changed to commenting through Facebook, meaning that HuffPo user profiles were *replaced* by Facebook profiles in a 'real name' environment. In this phase, comments appear below the line of the news article under the user's real name, as well as - depending on a user's privacy settings - appearing simultaneously on their Facebook page. This may have the effect that users comment on the HuffPo but speak to an audience located on Facebook. We read interpret this change as a move to more integrated and connected online identities.

While we found more durable online identities were associated with greater civility but less participation in online commenting, we also observed a shift in the sorts of issues on which users most frequently comment. This might be normatively encouraging: the relative rise in articles tagged Gay Voices, we speculate, may be a result of the inhibition of hostile and offensive commenters. Here greater civility seems to promote more participation and inclusive discussion. However, we also note a more general shift away from conflictual and politicized topics and towards 'safer' topics of celebrity, lifestyle and consumer issues. Turning to the second commenting change, we found that the Facebook phase exhibited markedly less argumentative engagement relative to the 'pseudonymous' phase, as evidenced both by an overall reduction in the proportion of replies, and by measures of deliberative quality (Fredheim et al., 2015a, 2015b). This finding points in the same direction as a recent Pew survey that found people were less likely to discuss a controversial issue on social media than in face to face conversation (Hampton et al., 2014).

With these findings in mind, we argue that the pseudonymous phase harnessed the beneficial effects of increased durability, without introducing the costs of connectedness. More broadly, we can link these dimensions of identification to particular democratic goods. The dimension of traceability seems to be associated with inclusion and exclusion, as users make a binary choice to stay on the platform. The dimension of durability or continuity of identity is associated with deliberative quality in so far as (theoretically) it exposes users to communicative accountability and (empirically) is associated with a higher density of reason-giving in commenting. The dimension of connectedness ties users to real-world relations. This favours civility, but it also promotes what Schudson calls 'sociable' conversation rather than more adversarial 'democratic' conversation (Schudson, 1997). We argue that durable and disconnected identities are more conducive to productive, issue-based debate between heterogeneous users. This bears on discussions of polarization. It also bears on the drive towards integration of social media and news discussion.

Bibliography

Fredheim, R., Moore, A. and Naughton, J. (2015a). Anonymity and Online Commenting: An Empirical Study. http://papers.ssrn.com/abstract=2591299 (accessed 27 November 2015).

Fredheim, R., Moore, A. and Naughton, J. (2015b). Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. *ACM Web Science 2015*. University of Oxford (accessed 27 November 2015).

Hampton, K., et al. (2014). Social Media and the 'Spiral of Silence' *Pew Research Center: Internet, Science and Tech* http://www.pewinternet.org/2014/08/26/social-media-and-the-spiral-of-silence/ (accessed 2 March 2016).

Marx, G. T. (1999). What's in a Name? Some Reflections on the Sociology of Anonymity. *The Information Society*, **15**(2): 99–112.

Nissenbaum, H. (1999). The meaning of anonymity in an information age. *The Information Society*, **15**(2): 141–44.

Resnick, P. and Friedman, E. J. (2001). The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, **10**(2): 173–99.

Ruesch, M. A. and Märker, O. (2012). Real Name Policy in E-Participation. *CeDEM 12 Conference for E-Democracy and Open Government 3-4 May 2012 Danube-University Krems, Austria*. Edition-Donau-Univ. Krems, pp. 109 (accessed 23 March 2015).

Indigenous digital humanities. Participatory geo-referenced-mapping and visualization for digital data management platforms in digital anthropology

Urte Undine Froemming
u.froemming@fu-berlin.de
Freie Universitaet Berlin, Germany

In this short paper presentation a participatory ethnographic GIS-mapping approach and an example for a digital platform for the visualization and digitalization of local environmental knowledge will be introduced. The term "Indigenous" is used to refer to those who "have a historical continuity with pre-invasion and pre-colonial societies that developed on their territories and consider themselves distinct from other sectors of the societies now prevailing in those territories" (UNESCO, 2004). Based on examples and research results of the research project ANIK *Alpine risks in times of climate change*, funded (2012-2015) by the German Ministry of Education and Research (BMBF), this paper will question how local environmental knowledge (Pottier, 2003) and local perceptions and handling strategies of climate-related risks may be gathered through participatory GIS mapping (PGIS), (Reichel, Frömming 2015, 2014). Based on applied visual anthropological methods, PGIS is a relatively new cartographical digital approach which includes local perceptions and strategies of action gathered during interviews and participant observation. This approach fosters active participation of local people (Wood, 2012), not only during the research process in the field, but also during the presentation process of the research data as digital spaces of memory.

Bibliography

Frömming, U. U., Reichel, Ch. (2014). Multimedia platform for geo-referenced data management http://medien.cedis.fu-berlin.de/cedis_medien/projekte/safiental/

Heather A., H., Miller, D. (eds.) (2012). *Digital Anthropology*. London: Berg Publishers.

Pottier, J. (2003). *Negotiating local knowledge: An introduction, Negotiating local knowledge: Power and identity in development*. In Pottier, J., Bicker, A. and Silitoe, P. (eds.), London: Pluto Press, pp. 1–29.

Reichel, Ch., Frömming, U. U. (2014). Participatory Mapping of Local Disaster Risk Reduction Knowledge. Example from Switzerland, *International Journal of Disaster Risk Science*, **5**(1): 41-54.

UNESCO - United Nations Declaration on the Rights of Indigenous People (2004) and (2006). Publisher: United Nations.

Wood, D. (2012). The anthropology of cartography. In Roberts, L. (ed), *Mapping cultures: Place, practice, performance*. New York: Palgrave Macmillan, pp. 280-303.

Standardized Digital workflow for Archiving Local Knowledge

Yu Fujimoto
yfujimot@daibutsu.nara-u.ac.jp
Nara University, Japan

Yasuhiko Horiuchi
horiuchi.yasuhiko@gmail.com
NPO “The Field”, Japan

Preface

The declining birth rate and resulting increase in the proportion of the elderly are serious issues for contemporary Japan. The National Institute of Population and Social Security Research reported that the 2010 Japanese population of about 128,057,000 is expected to decline to about 86,737,000 in 2060 with depopulation accelerating most acutely in local regions (The National Institute of Population and Social Security Research, 2014). Currently, in 7,878 villages in Japan, over half of the population is over the age of 65. From this figure, it is predicted that 2,500 villages will vanish in the next 10 to 30 years (Ministry of Land, Infrastructure, Transport and Tourism, 2014; Masuda, 2014). Given this situation, the preservation of local knowledge is essential and standardized digital archiving methods are required.

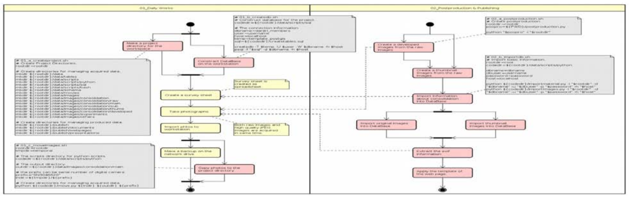

01_Daily Works
02_Postproduction & Publishing
Make a project directory for the workspace
Construct DataBase on the workstation
Create a survey sheet
Take photographs
Import photos to workstation
Make a backup on the network drive
Copy photos to the project directory
Create a developed images from the raw images
Create a thumbnail images from the raw images
Import information about consolidation into DataBase
Import original images into DataBase
Import thumbnail images into DataBase
Extract the exif information
Apply the template of the web page.

Figure 1: The workflow using MILC for digital archiving

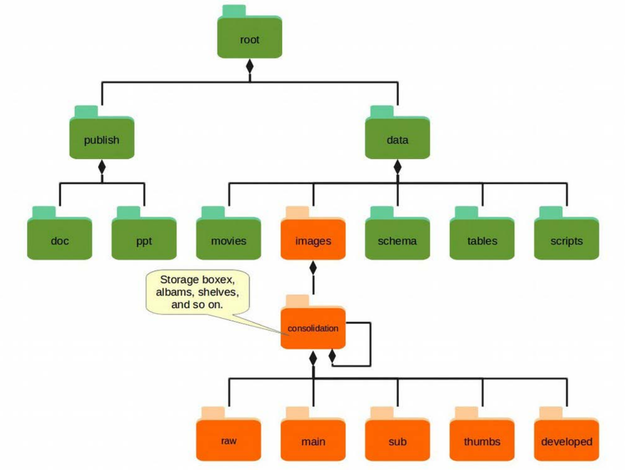


Figure 2: Directory structure for acquired data

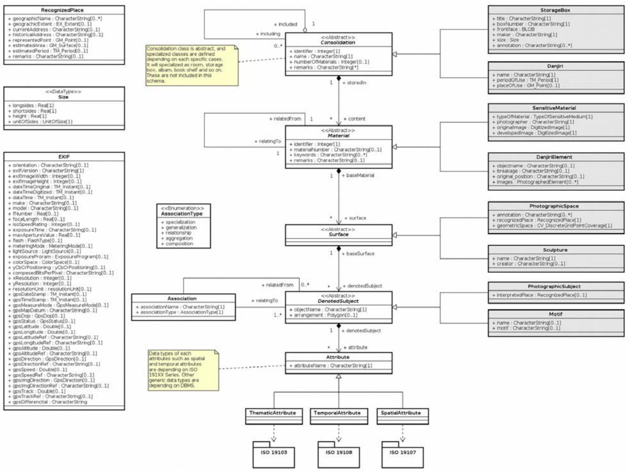

RecognizedPlace
+ geographicName : CharacterString[0..*]
+ geograchicExtent : EX_Extent[0..1]
+ currentAddress : CharacterString[0..1]
+ historicalAddress : CharacterString[0..1]
+ representedPoint : GM_Point[0..1]
+ estimatedArea : GM_Surface[0..1]
+ estimatedPeriod : TM_Period[0..1]
+ remarks : CharacterString[0..1]
<<DataType>> Size
+ longsides : Real[1]
+ shortsides : Real[1]
+ height : Real[1]
+ unitOfSides : UnitOfSize[1]
EXIF
+ orientation : CharacterString[0..1]
+ exifVersion : CharacterString[1]
+ exifImageWidth : Integer[0..1]
+ exifImageHeight : Integer[0..1]
+ dateTimeOriginal : TM_Instant[0..1]
+ dateTimeDigitized : TM_Instant[0..1]
+ dataTime : TM_Instant[0..1]
+ make : CharacterString[0..1]
+ model : CharacterString[0..1]
+ fNumber : Real[0..1]
+ focalLength : Real[0..1]
+ isoSpeedRating : Integer[0..1]
+ exposureTime : CharacterString[0..1]
+ maxApertureValue : Real[0..1]
+ flash : FlashType[0..1]
+ meteringMode : MeteringMode[0..1]
+ lightSource : LightSource[0..1]
+ exposureProram : ExposureProgram[0..1]
+ colorSpace : ColorSpace[0..1]
+ yCbCrPositioning : yCbCrPositioning[0..1]
+ compesedBitsPerPixel : CharacterString[0..1]
+ xResolution : Integer[0..1]
+ yResolution : Integer[0..1]
+ resolutionUnit : resolutionUnit[0..1]
+ gpsDateStamp : TM_Instant[0..1]
+ gpsTimeStamp : TM_Instant[0..1]
+ gpsMeasureMode : GpsMeasureMode[0..1]
+ gpsMapDatum : CharacterString[0..1]
+ gpsDop : GpsDop[0..1]
+ gpsStatus : GpsStatus[0..1]
+ gpsLatitude : Double[0..1]
+ gpsLongitude : Double[0..1]
+ gpsLatitudeRef : CharacterString[0..1]
+ gpsLongitudeRef : CharacterString[0..1]
+ gpsAltitude : Double[0..1]
+ gpsAltitudeRef : CharacterString[0..1]
+ gpsDirection : GpsDirection[0..1]
+ gpsDirectionRef : CharacterString[0..1]
+ gpsSpeed : Double[0..1]
+ gpsSpeedRef : CharacterString[0..1]
+ gpsImgDirection : GpsDirection[0..1]
+ gpsImgDirectionRef : CharacterString[0..1]
+ gpsTrack : Double[0..1]
+ gpsTrackRef : CharacterString[0..1]
+ gpsDifferencial : CharacterString
Consolidation class is abstract, and specialized classes are defined depending on each specific cases. It will specialized as room, storage box, albam, book shelf and so on. These are not included in this schema.
+ included
1
+ including
0..*
<<Abstract>> Consolidation
+ identifier : Integer[1]
+ name : CharacterString[1]
+ numberOfMaterials : Integer[0..1]
+ remarks : CharacterString[*]
1
+ storedIn
+ relatedFrom
1
*
+ content
+ relatingTo
1
<<Abstract>> Material
+ identifier : Integer[1]
+ materialNumber : CharacterString[0..1]
+ keywords : CharacterString[0..*]
+ remarks : CharacterString[0..1]
1
+ baseMaterial
<<Enumeration>> AssociationType
+ specialization
+ generalization
+ relationship
+ aggregation
+ composition
*
+ surface
<<Abstract>> Surface
1
+ baseSurface
+ relatedFrom
0..*
*
+ denotedSubject
Association
+ associationName : CharacterString[1]
+ associationType : AssociationType[1]
+ relatingTo
1..*
<<Abstract>> DenotedSubject
+ objectName : CharacterString[1]
+ arrangement : Polygon[0..1]
1
+ denotedSubject
*
+ attribute
Data types of each attributes such as spatial and temporal attributes are depending on ISO 191XX Series. Other generic data types are depending on DBMS.
Attribute
+ attributeName : CharacterString[1]
ThematicAttribute
TemporalAttribute
SpatialAttribute
ISO 19103
ISO 19108
ISO 19107
StorageBox
+ title : CharacterString[1]
+ boxNumber : CharacterString[1]
+ frontface : BLOB
+ maker : CharacterString[1]
+ size : Size
+ annotation : CharacterString[0..*]
Danjiri
+ name : CharacterString[1]
+ periodOfUse : TM_Period[1]
+ placeOfUse : GM_Point[0..1]
SensitiveMaterial
+ typeOfMaterial : TypeOfSensitiveMedium[1]
+ photographer : CharacterString[1]
+ originalImage : DigitizedImage[1]
+ developedImage : DigitizedImage[1]
DanjiriElement
+ objectname : CharacterString[0..1]
+ breakage : CharacterString[0..1]
+ original_position : CharacterString[0..1]
+ images : PhotographedElement[0..*]
PhotographicSpace
+ annotation : CharacterString[0..*]
+ recognizedPlace : RecognizedPlace[1]
+ geometricSpace : CV_DiscreteGridPointCoverage[1]
Sculpture
+ name : CharacterString[1]
+ creator : CharacterString[0..1]
PhotographicSubject
+ interpretedPlace : RecognizedPlace[0..1]
Motif
+ name : CharacterString[0..1]
+ motif : CharacterString[0..1]

Figure 3: Application Schema for Local Knowledge (ver.2015-10-11)

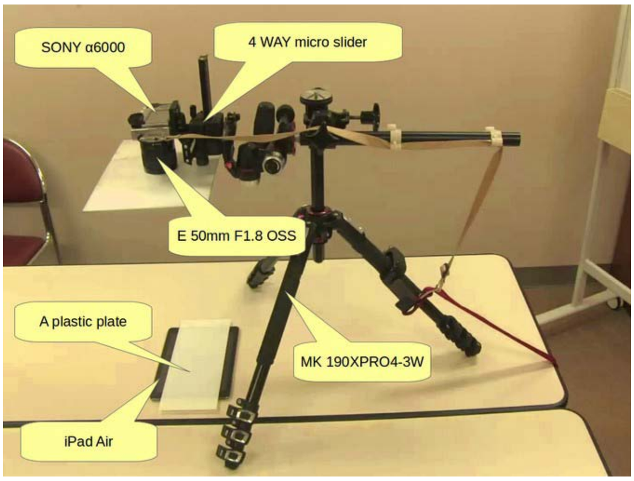


Figure 4: Digital Archiving of photographic dry plates

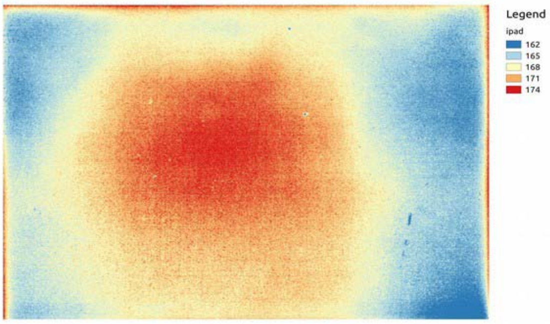


Figure 5: Brightness distribution of iPad Air

In contrast to national treasures, digital archiving for local cultural properties is limited in various respects. Digitizing devices should be versatile because the kind of materials buried is unknown at the onset of the survey. Budgets are therefore limited because most people in possession of such items do not have sufficient financial background. Therefore, the methods for archiving local items and knowledge should be as low-cost as possible. In addition to these problems, a standardized digital archiving workflow and information management rules are required to perform distributed autonomous digital archiving activities. Formulating digital archiving rules that are both open and standardized is also essential in terms of the Long Term Preservation issue (Lorie, R.A., 2001).

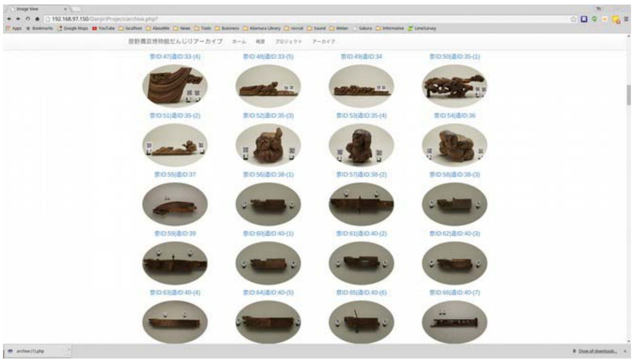


Figure 6: The resulting digital archive system

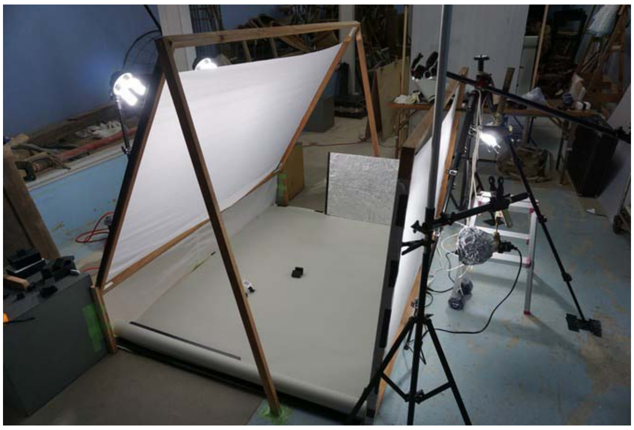


Figure 7: Handmade photo studio for Danjiri elements

Responding to these issues, this paper will propose a low-cost digitizing workflow using an Mirror-less Interchangeable Lens Camera (MILC) and the information standard for local memories and knowledge with the concept of a Work-Oriented Approach (WOA) (Fujimoto 2011). The proposed methods have already been used in two different types of experimental projects and the results will also be summarized in this paper.

The digital archiving workflow

The digital camera is excellent in terms of saving space, versatility, speed of digitizing and later correction, and is therefore the ideal device for basic digital archiving. In particular, the Mirror-less Interchangeable Lens Camera (MILC) has a rich choice of lenses, and is lighter and therefore preferable to than the general Digital Single Lens Reflex camera (DSLR).

Figure 1 shows the workflow of constructing a digital archive using the MILC. In this figure, the workflows are denoted with UML activity diagrams, with automated activities in red, and yellow designating those activities conducted manually. Each automated activity is described by BASH scripts to ensure the clarity of specific procedures, and some call python scripts utilizing open source libraries. In practice, Graphical User Interfaces (GUI) should be provided for ordinal users.

This activity diagram is separated into two lanes, with the left lane showing the digital archiving workflow and the right lane representing the post-production and publishing workflow. In the digital archiving workflow, taking photographs, writing investigative reports, and transferring acquired images to the work station are manually conducted, whereas other activities are done automatically.

To perform these automatic activities, a working directory should be defined as shown in Figure 2. The raw images acquired are automatically copied to the "*raw*" directory and JPEG format images are initially copied to "*main*", and then the raw images are developed and saved in the "*developed*" directory. Thumbnails are saved in "*thumbs*". Finally, unused images are moved to "*sub*" directories. By defining the directory structure, all projects are generalized, and creating an automatic workflow becomes much easier.

The application schema

Promoting such shared and mutual use is essential and will help preserve local memories and knowledge for future generations. To do so, it is important to conform to conventional international standards (Fujimoto, 2009). The ISO 191XX series is a versatile information standard for database construction. This international standard is based on the idea of object-oriented GIS, and it defines a meta-model of geographic features and spatio-time objects. The ISO 191XX series is generally adopted for public geographic information suchas census or infrastructure data. However, it has many other possible applications. Additionally, this standard provides encoding rules by XML, and metadata, geometric information, raster format datasets and tabular form attributes can be denoted as XML elements. Therefore, this standard is also effective in terms of Long Term Preservation issues.

The Figure 3 is the application schema proposed in this paper, which enables the storage of unknown tangible as well as intangible cultural properties in compliance with the standard. This application schema is based on the proposed working flow, and weighs heavily in extendibility. The fundamental classes including the *Consolidation* class, *Material* class, *Surface* class and *DenotedSubject* class are *Abstract* classes, which are actually implemented in specific ways. The grey-colored classes are examples applicable to photographic dry plates and " *Danjiri*" elements. The attributes for these classes are minimum essential attributes, and attributes relevant to each specific item can be defined using the Attribute class. This class specializes three classes: ThematicAttribute, SpatialAttribute and

TemporalAttribute classes. Any kind of attribute can be defined using this classification.

Case studies

The proposed standardized workflow and the database schema have been used in two case studies. One is the digital archiving project of dry photographic plates taken between the *Meiji* period and mid-*Showa* period, while the other is a "*Danjiri*", a traditional large wooden cart used for traditional Japanese festivals, which was completely destroyed by a flood.

In the first case a tablet device, iPad Air, was used as a substitution for a light box, and SONY A6000 was used as a digitizing device (Figure 4 and Figure 5) (Fujimoto, 2015). By using a set of these ordinary devices with three workers, more than three hundred old dry plates were duplicated in about 13 hours in total. In this project, all of the operations were composed almost entirely of open-source software, and performed in batches. Although this project was successfully completed, some improvements were later found for making linkages between the acquired images and investigative reports and application schema.

In the latter study, the digital archiving workflow and the application schema, originally designed for photographic dry plates, were modified for versatility, to make them applicable to the Danjiri elements. In this project, more than three hundred fragments of Danjiri elements were archived in one week by four workers with one day required for setting up the photographic studio, four days for taking photos, and two days for developing the digital archive system (Figure 6). In contrast to the former project, the digital archiving subjects were three-dimensional and lighting instruments were required. To achieve better results within a limited budget, a lighting studio was built using wooden blocks and domestic fluorescent lamps, which were procured on the site (Figure 7).

Conclusions

In digital archiving projects focusing on important cultural properties, intended objects are well known and professional and/or specialized equipment are used to achieve with the best possible results. However, to construct digital archives of locally-kept cultural properties, versatile and inexpensive methods are required. Additionally, a standardized workflow and database schema covering various kinds of materials should be considered.

In this paper, using the concept of WOA, a reasonable MILC is used as a digitizing device, and a standardized workflow enabling automation and database schema compliant with ISO 191XX are proposed. Because the proposed workflow, information schema and libraries utilize the existing international standard and open-source technologies, outcomes including metadata and source codes can be opened. This method is therefore effective in terms of the Long Term Preservation issues. These standardized methods have been tried in two different types of experimental case studies. Although both projects were successfully completed, some continual refinement is necessary to perform a fully automated workflow, especially post-production.

Finally, the methods proposed in this paper can be also applicable in disaster restoration. It is important to preserve memorial items of ordinary people such as family photo albums in order to incite the energy vital for the restoration process. Unfortunately, in fact, after the Great East Japan Earthquake in 2011, a large number of mementoes and keepsakes of the local people were discarded. In such cases low-cost, swift and standardized digital archiving methods are essential.

Bibliography

Fujimoto, Y. (2009). Information Standards for Cultural Heritage with The ISO 191XX Series, *Proceedings of the 22nd CIPA Symposium*, Kyoto, Japan.

Fujimoto, Y. and Horiuchi, Y. (2011). Using Standards as a Bridge Between Traditional Research and Technologies in Protecting Cultural Assets, *Proceedings of the 23rd CIPA Symposium*, Prague, Czech Republic.

Fujimoto, Y. (2015). Digitizing Dry Plates of the Kitamura Collection in Nara University Library, *Memoirs of Nara University*, No. 43, pp. 91-102. (Japanese)

Lorie, R., A. (2001). Long Term Preservation of Digital Information. *Proceedings of the first AM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, Virginia, New York: ACM-Press, pp. 346-52.

Ministry of Land, Infrastructure, Transport and Tourism. (2006). *Assessment of Situation with Settlements to Institute National Spatial Planning*. (Japanese)

Masuda, H. (2014). The Death of Regional Cities: A horrendous simulation Regional Cities Will Disappear by 2040 A polarized Society will Emerge, *Discuss Japan – Japan Foreign Policy Forum*, 18.

The National Institute of Population and Social Security Research (2014). *Estimation of the Japanese Future Population* (Japanese).

Playing with French Drama: from Old Research Questions to New Research Tools

Ioana Galleron
ioana.galleron@univ-ubs.fr
LIDILEM, U. Grenoble Alpes, France

While early English drama has been subject to several DH investigations, because "a quite explicit and rigid system of metadata is part of the genre itself" (Mueller, 2014: par27), French theatre of the 17th and 18th century remains largely underexplored, in spite of demonstrations and case studies presented by authors with large international audiences (Moretti, 2005). This paper presents some results of a project inspired by the explorations of the English corpora, but dedicated to the computer aided identification and analysis of theatrical topoi in the French domain.

Topoi can be defined as "recurrent configurations of thematic or formal elements" (adapted from Weill and Rodriguez, SATOR, 1996) and they are highly significant for the French plays, which are largely based on a common stock of characters, scenes and themes, and in which novelty often consists in reuse with a shift of former materials. Considering the vast amount of plays produced over the period going from, roughly, 1630 to 1790 (some 12000 plays, even if the text of many of them is lost), the use of computational tools to track the presence of various topoi over the time appears appropriate. Computer aid can help answer in a (more) rigorous way questions concerning the aesthetics of the French theatre of the period, too often reduced to "classicism", and historical interrogations about moments and reasons when certain meaning units are more in use or relatively abandoned.

While the results of this endeavour are at a very early phase, for reasons to be detailed in the presentation, the very attempt to clarify some of these literary questions shows that the use of the computer is not neutral and modifies the problems to be solved. After a more detailed presentation of the activities undertaken to date, this paper will reflect upon the way in which the methods and tools used to approach the texts modify the initial research questions.

Dramacode: an expanding community for French drama

Identifying topoi using a computer asks in the first place for the existence of digitised texts in a more elaborate format than the pdfs or plain texts offered by Gallica (see http://gallica.bnf.fr). Various initiatives have been taken in this domain in France, either by scholars supported by HEIs and research institutes (see "Molière project" http://obvil.paris-sorbonne.fr/corpus/moliere/moliere) or by individuals with computer skills and a special interest in literature from the 17th and 18th century (see http://www.theatre-classique.fr). Unfortunately, this diversity of initiatives resulted in a great variety of encoding practices, more or less TEI compliant. To reduce this diversity and move towards common standards, an organisation has been created on GitHub. Under the name of "Dramacode", it hosts some 882 plays written or staged between 1630 and 1810 (roughly, 7,3% of the dramatic production of the period), to be progressively re-tagged and enriched with further mark-up. "Dramacode" allows also to publish the contents with a harmonised style-sheet, and to share work on experiments of visualisation of data extracted from the plays.

Finding theatrical topoi through linguistic analysis and literary mark-up

Within this corpus, the plays of Louis de Boissy (1694-1754) have been selected for a pilot study, destined to build strategies for gathering sufficient clues, through various queries, about the existence of a topos in an unknown (i. e., not previously read) play, digitised and encoded in another project.

Louis de Boissy was a French play writer who enjoyed a certain success in the 18th century mainly because of two of his plays, *Le Français à Londres* and *L'Homme du jour*. Beyond these achievements, what renders his work interesting is precisely its mediocrity. Boissy's works are a perfect example of the reuse of topoi, sometimes to follow the fashion, sometimes in an attempt to distinguish himself by revitalising conventional themes and dialogues. He has also the particularity of having worked for the three main theatrical spaces of the period, namely the Comédie-Française, Théâtre Italien and Théâtres de la Foire. Therefore, his texts become a privileged observatory of French comedy in the first half of the 18th century.

Two approaches of topoï identification are developed. The first one consists in developing an ontology for characters and scenes description. Roles and description of roles from the existing plays have been extracted, cleaned, and analysed using an excel spread-sheet. Ten categories appear necessary for describing characters from classical French scenes; five of them were known and formalised since Aristotle, or, at least, since the rediscovery of Aristotle by French authors and thinkers of Renaissance, then 17th century: nationality ("climat" in Marmontel's words, see entry "Moeurs" in Eléments de littérature, 1787), age, sex, occupation ("état") and temperament. Five others were deemed of interest, as allowing potential automatic analysis aimed to spot changes in aesthetic principles: the ontological category (human/ nonhuman/ semihuman), the dramatic or actantial role, the social status ("king", "prince", "marquis", "Dom"…: it seemed useful to distinguish these from information about “état”), the family position, and the collective nature of some characters.

In parallel, types of scenes and other significant recurrences are manually identified and marked-up as “recognition”, “fight”, “complot”, etc. Relevant lines or phrases (labelled with xml:ids) are regrouped as <span>s in the <back> section of <text>; if each span receives, for the moment, a literal description, this will progressively feed into another library and will be later referred to using @ana.

The second approach is based on a linguistic analysis of the speeches, using AntConc to generate concordances and to identify key-words. The initial objective was to spot key-words or linguistic idiosyncrasies of the character of “petit-maître”, whose identification was the primary goal of this research. Indeed, “petit-maître” are not always labelled as such, and therefore their presence in French drama has been underestimated and incorrectly related to a specific period in time (mainly, “decadence comedy” after 1680). If, in time, a complete mark-up of the characters according to the above-mentioned categories (i. e., <temper>) will help identifying as “petits-maîtres” those who are not called as such by roleDesc or other characters in the play, speech identification is for the moment the only option for answering the research question.

New research vistas

When observing the trajectory of this study, the digital approach can be said as having significantly modified the initial perspective and research questions. The TEI tagging of characters and scenes revealed the need to refine methods and categories developed by the narratology and the theatrical analysis in 1970 and 1980, preparing the grounds for more accurate description of character building.

In the meantime, the effort to describe a stylistic specificity of the “petits-maîtres” revealed to be rooted in a larger, two folded research question, concerning the potential differences in speech between characters of the same author. Digital humanities have been, to date, more interested in global stylistic analysis on one side, and in authorship studies, on the other side, while this research project intends to apply the methods of the second approach to fictional speakers. While such an endeavour can be justified by the pretences of French dramatists of the classical centuries to act as kind of “sociolinguists”, the question is as well that of the methods to automatize the analysis of specific speeches when conducted on a large scale (more than 9300 distinct characters presents in the above-mentioned set of plays), and of the explanations to be found for the observed recurrences. Indeed, during the pilot study executed on Boissy plays, it appeared that his characters group, by their speech, less in male and female speakers, but rather in two unexpected classes, one formed by those who tend to speak more about themselves (characters saying “je”), the second by those who address themselves to the the others (characters saying “vous”). The importance of “I” and “You” has already been spotted in theatrical studies (see Craig, 2004), but not the observed repartition, which needs, however, to be checked against the larger corpus, a very time consuming operation for the moment.

It is to be expected that further observation of concordancers and n-grams extracted from very large collections of digitised plays will put into light such other unexpected recurrences, triggering a new understanding about what is a classical play, and contributing to the on-going debate about the nature of the theatre.

Bibliography

Marmontel, J. F. (2005). *Eléments de littérature. Édition présentée, établie et annotée par Sophie Le Ménahèze*. Paris : Desjonquères.

Mueller, M. (2014). Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment. *DHQ: Digital Humanities Quaterly*, **8**(3).

Moretti, F. (2005). *Graphs, maps and trees: Abstract Models for a Literary Theory*. London and New York: Verso.

Craig, H. (2004). Stylistic Analysis and Authorship Studies. In Schreibman, S. Siemens, R. and Unsworth, J. (Eds.) *A Companion Digital Humanities*. Oxford: Blackwell, DOI: 10.1002/9780470999875.ch20.

Big Data and the Study of Allusion: an Exploration of Tesserae's Multitext Capability

James O'Brien Gawley
jamesgaw@buffalo.edu
University at Buffalo, United States of America

A. Caitlin Diddams
acstaab@buffalo.edu
University at Buffalo, United States of America

This paper aims to explore the methodology and the potential of the “multi-text tool,” an advanced feature of the Tesserae Project. This tool is capable of identifying potentially interesting textual parallels between works, and collating all instances of these and similar phrases throughout the entire corpus of canonical Latin or Greek. As such it expands the researcher's ability to identify meaningful intertexts among Tesserae results, and makes it possible to quantify the literary influences which act upon a given work.

Tesserae is an online tool that aims to automatically identify allusions and more general forms of intertext between ancient authors. The program can identify specific intertexts between two works, as in previous research which has brought to light new parallels between Lucan's *Bellum Civile* and Vergil's *Aeneid*. A standard Tesserae search identifies a possible intertext when the same words (regardless of inflection) appear within the same phrase in two different texts. This 'bigram' measurement of similarity has been demonstrated to capture roughly 67% of intertexts previously noted in scholarship. The multitext function begins with a standard two-text Tesserae comparison; all bigrams shared between the two texts are then compared to a corpus of additional works selected by the user.

Although the first aim of most Tesserae users is to discover previously unnoted allusions, not all bigram similarities discovered by a Tesserae search are allusions. The program's scoring algorithm attempts to sort the meaningful intertexts sought by the researcher from undesired, coincidental overlap by considering the rarity of the shared words and their proximity to one another. This method has shown to be at least partially effective, yet it does not fully predict the assesment of the expert researcher. Because Tesserae results can number in the tens of thousands, further means of identifying desirable intertexts is necessary, particularly when the number of results is expanded by new search features such as semantic matching, introduced in October 2015. Tesserae's multitext search can be used to trace the history of a phrase throughout the corpus, and thereby eliminate oft-repeated bigrams. This method assumes that phrases appearing in many previous works are less likely to represent allusions.

In addition to identifying new allusions between two works, intertextual scholars often wish to measure the rate of connection between them. The level of active engagement between authors suggests the literary influence of a given work, and we propose a way to quanitfy this engagement. Using the multitext tool, the researcher can eliminate all widely-occuring textual parallels between potentially connected works and retain only unique results, then consider the rate of unique intertextuality against a baseline figure. Unlike the micro-level analysis of individual textual parallels, this macro-level analysis allows the reader to examine the intertextuality between works as a whole.

Our preliminary research shows that specific textual parallels with a large number of multitext results tend to indicate the use of generalized language rather than a meaningful communication between works. For example, Augustine's language in *De Doctrina Christiana* bears measureable resemblance to the language of Quintilian, Cicero, and Tacitus. Yet scholars have long argued that Augustine's primary influence was Cicero. Our multitexts results show Augustine's engagement with Cicero include a large percentage of "unique" intertexts, seldom picked up by other authors. His connections to other authors were mostly composed of "general" intertexts, which appear to consist of standard language used by almost all Latin writers engaged in the discussion of rhetoric. The same method has been previously used to examine Claudian's differential level of interaction with the various authors of Latin epic. The examples of Claudian and Augustine demonstrate the efficacy of unique intertextual connections as a measurement of relative literary engagement.

Although the elimination of oft-repeated language is effective for eliminating meaningless intertexts when considering works at the macro-scale, the scholar in pursuit of specific, meaningful intertexts should not eliminate these parallels. Sometimes a phrase with a large number of multi-texts results is actually more significant than one with fewer multi-text results. Certain phrases are quoted and alluded to so often that they become "viral." These instances do tend to apply in cases where a Tesserae match is ranked particularly high, but it ultimately remains the responsibility of the reader to sort meaningful intertexts from coincidental ones.

The multi-text tool expands the potential big-data research in the field of Classics, allowing for quantitative stylistic analysis at a rate of speed and a level of specificity previously impossible. It is an essential addition to the basic matching of the Tesserae Project, and we hope to see it become an element of digital humanities programs in various academic curricula. From specific analyses of sets of texts to more comprehensive explorations of style, the multi-text tool should be a fundamental resource in the study of allusion in Latin literature. In addition to demonstrating the efficacy of this tool, in this paper we explain how to use the results generated by a multi-text search to build an empirical measurement of authorial engagement, beginning with how to run the module on a personal computer and concluding with how to meaningfully analyze its results and tailor its parameters for individual projects.

Bibliography

Coffee, N. et al. (2012). The Tesserae Project: Intertextual Analysis of Latin Poetry. *Literary and Linguistic Computing*, **28**(1): 221-28. doi: 10.1093/llc/fqs033.

Coffee, N. et al. Modeling the Interpretation of Literary Allusion with Machine Learning Techniques. *Journal of Digital Humanities*, **3**(1).

Coffee, N. and Forstall, C. (forthcoming). Claudian's Engagement with Lucan in his Historical and Mythological Hexameters. In Berlincourt, V., Galli-Milic, L. and Nelis, D. (eds), *Lucain et Claudien face à face: une poésie politique entre épopée, histoire et panégyrique*. Winter Verlag.

Conley, T. M. (1990). *Rhetoric in the European Tradition*. New York: Longman.

Murphy, J. J. (1960). Saint Augustine and the Debate about a Christian Rhetoric. In Enos, R. L. and Thompson, R. (ed.), *The Rhetoric of St. Augustine of Hippo: De Doctrina Christiana and the Search for a Distinctly Christian Rhetoric*. Waco: Baylor University Press, pp. 205-17.

Forstall, C. et al. (2015). Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching. *Literary and Linguistic Computing*, doi: 10.1093/llc/fqu014.

Gawley, J., Forstall, C. and Coffee, N. Evaluating the literary significance of text re-use in Latin poetry, *Poster presented at Chicago Colloquium on Digital Humanities and Computer Science*, University of Chicago, Chicago, IL. November, pp. 17-19.

Biblissima - Following Medieval Manuscripts and Incunabula through their Existence via a Semantic Web Application

Stefanie Gehrke
stefanie.gehrke@biblissima-condorcet.fr
Equipex Biblissima, France

Biblissima—*Bibliotheca bibliothecarum novissima*—is a Digital Humanities project that brings together libraries and research institutions, whose goal is to provide a single access point to over 40 databases regarding medieval and Renaissance manuscripts and early printed books by the end of 2016.

The project began late in 2012 with nine founding partners: the Department of Manuscripts at the BnF (Bibliothèque nationale de France), the Campus Condorcet, the CESR (Centre d'Etudes Supérieures de la Renaissance), the CIHAM (Histoire, Archéologie, Littératures des mondes chrétiens et musulmans médiévaux), the Centre Jean-Mabillon at the ENC (École nationale des chartes), the CRAHAM (Centre de Recherches Archéologiques et Historiques Anciennes et Médiévales. Centre Michel de Boüard), the IRHT (Institut de recherche et d'histoire des textes), the Digital Document Centre at the MRSH de Caen (Maison de la Recherche en Sciences Humaines), and the SAPRAT-EPHE (Savoirs et pratiques du Moyen Âge au XIXe siècle, Ecole pratique des hautes études).

Funded by the French National Research Agency (ANR: Agence Nationale de la Recherche), Biblissima is concerned with the history of collections and the transmission of texts. In this respect, the project stands somewhat in line with the works of Antoon Sanders, who published his *Bibliotheca Belgica Manuscripta* in 1641-44, and Bernard de Montfaucon and his *Bibliotheca bibliothecarum manuscriptorum nova*, published in 1739, both of which are major inventories compiling lists of manuscripts held in many different libraries in Europe.

In order to create an *Online library of historical collections of France for the 21st century*, our chosen solution was to develop a semantic web application. We released a prototype (http://demos.biblissima-condorcet.fr/prototype) in summer 2015 that provides unified access to a subset of two major iconographic databases: Initiale (http://initiale.irht.cnrs.fr/accueil/index.php) and Mandragore (http://mandragore.bnf.fr/html/accueil.html). An export of the metadata pertaining to the illuminations that depict geographical locations was used as a starting point for the development of our application. People (author or illuminator), institutions (libraries), work titles and places were aligned with the BnF authority records and, whenever possible, also linked to external vocabularies (GeoNames, VIAF, Pleiades, Getty thesauri).

The next step involved generating dynamic web pages that describe a person, a work, an expression, a manuscript, a part of a manuscript or an illumination, and which include links to the corresponding descriptions in the source databases. This original data is enriched with longitude and latitude coordinates for geographical names that were acquired by aligning them with the equivalent GeoNames concepts. It is now possible, for example, to show on two different maps the places depicted in illuminations with the images of the corresponding folios, and the illuminations from manuscripts held in a particular institution.

In addition, a viewer embedded in the page (we currently favour Mirador: https://iiif.github.io/mirador) shows the digitised manuscript or folio when available. The information needed to display these images is structured according to the Shared Canvas data model (http://iiif.io/model/shared-canvas), which is the foundation of IIIF (International Image Interoperability Framework: http://iiif.io). This information is passed to the viewer in form of JSON-LD manifests, which are generated from metadata and image files supplied by the partner institutions. The viewer itself is client-based and features a deep-zoom capability for the loaded images, the possibility to display associated metadata, and also the superposition of different layers, such as an image and the corresponding textual transcription.

The data model behind the application is based on CIDOC-CRM (http://www.cidoc-crm.org) and FRBRoo (http://www.cidoc-crm.org/frbr_inro.html) and our data is already available in RDF format. Beyond taking into account the different levels of work, manifestation and item (using the classes F4 Manifestation Singleton for manuscripts as well as F3 Manifestation Product Type and F5 Item for early printed books), we intend to group manuscripts and early printed books as productions of certain expressions (class F2 Expression). Illuminations are understood as features (class E26 Physical Feature) placed on a folio or page, and provenance marks will be modelled in the same way. Based on the lessons learned through developing our own ontology, we will be able to give feedback about the data.bnf.fr, FRBRoo and SharedCanvas ontologies, as well as other theoretical foundations.

We are currently working on extending this prototype, which was developed using the semantic web application framework CubicWeb (https://www.cubicweb.org/). In addition to manuscripts, we will be including early printed books, with a strong emphasis on book provenance, by integrating data from the following four sources: Esprit des Livres (database on auction and other sales catalogues, ENC), Bibale (database on historical book collections, IRHT), Europeana Regia (database on three important historical collections of the Middle Ages and Renaissance), as well as CR2I (Catalogues Régionaux des Incunables Informatisés, CESR) and CRIICO (CR2I Centre-Ouest, CESR). Resolved challenges include the creation of ARKs (Archival Resource Keys) for each manuscript (BnF and IRHT) and codicological unit (BnF). At a later stage, digital editions of inventories of manuscripts in TEI XML will be integrated as well.

This short paper will present the latest version of the semantic web application for Biblissima's data cluster.

DH Poetry Modelling: a Quest for Philological and Technical Standardization

Elena González-Blanco
egonzalezblanco@flog.uned.es
Universidad Nacional de Educación a Distancia (UNED), Spain

Gimena Del Rio Riande
gdelrio@conicet.gov.ar
Seminario de Edición y Crítica Textual (SECRIT-IIBICRIT CONICET), Argentina

Clara Martínez Cantón
Cimartinez@flog.uned.es
Universidad Nacional de Educación a Distancia (UNED), Spain

Introduction

Standardization has become an increasingly important process in relation to academic research, as it provides a better way for exchanging information. Humanities and cultural studies have followed, however, a heterogeneous path in which creativity and tradition play an essential role. Comparative studies in literature, and especially poetry, are a clear example of this eclectic situation. There is not a uniform academic approach to analyze, classify or study the different poetic manifestations. Things get even worse when comparing poetry schools from different languages and periods. The result of this uncoordinated evolution is a bunch of varied terminologies that means to explain analogous metrical phenomena through the different poetic systems whose correspondences have been hardly studied (González-Blanco and Sélaf, 2014).

The existing digital poetic repertoires and databases are a good example of this situation, as they constitute a rich kaleidoscope of multilingual virtual poetry, constituted by lyrical collections in French (*Nouveau Naetebus*), Italian (*BedT*), Hungarian (*RPHA*), Medieval Latin (*Corpus Rhythmorum Musicum, Annalecta Hymnica Digitalia, Pedecerto*), Gallego-portuguese (*Oxford Cantigas de Santa María, MedDB2*), Castilian (*ReMetCa*), Dutch (*Dutch Song Database*), Occitan (*BedT, Poèsie Neotroubadouresque, The last song of the Troubadours*), Catalan (*Repertori d'obres en vers*), Skaldic (*Skaldic Project*), or German repertoires (*Lyrik des Minnesänger*), among others.

Interoperability among these different poetic corpora would be desirable, as having a common search engine to extract information from all of these databases at the same time would have a deep impact for comparative studies in literature, linguistics and other humanities disciplines. We are, however, far from this reality as interoperability is not possible due to a lack of standardization both in technological and philological fields.

Philological standardization

During the Middle Ages and the Renaissance, the powerful influence of Latin made scholars inherit the terminology of Classical poetry treatises and apply it to Romance languages, regardless of their different linguistic traits and verse structures. When vernacular theories started to arise, each literary school set up its own terminology and classification system. This multiplicity led to complex situations, such as the creation of conceptual genres that only exist in some traditions.

Spanish conceptualization models are a good example to illustrate this situation: the classical system of Bello (1955), first published in 1835, divided all the structures into binary and ternary feet (imitating the classical Latin terminology):

Binary		**Ternary**	
Trochaic:	óo	Iambic:	oó
Dactylic:	óoo	Amphibrachic:	oóo
		Anapestic:	ooó

Later, the musical analysis system applied by Navarro Tomás (1956) was followed as a valid system through many years, using concepts like *anacrusis*. In the last years, there have been many different approaches to explain the Spanish panorama, as it is shown by the semantic comparative model designed by the Czech Oldrich Bělič.

The international context is richer, especially in English, with two prominent schools: 1) A traditional approach based on stress and classical feet; and 2) A generative approach based on the terminology and concepts shown through text grids that take into account word boundaries, with a strong impact on poetic theories (Gerber, 2013: 147).

The models described are just an example of the idiosyncrasy that can be observed in each literary tradition. Although the current ICT infrastructures are prepared to harvest different types of collections and models, it is necessary at a first stage to standardize metadata and map vocabularies and terminology at the philological level in order to build a consistent able to be shared between the different traditions.

Technological standardization

The lack of unified criteria is translated into many different uncoordinated technologies when research data are transformed to build digital projects and do not even follow a standard, in most cases. The multiplicity of technologies used includes SQL databases, TEI and XML markup languages, semantic web technology standards (RDF, OWL, SKOS), natural language processing systems (NLP) and visualization tools.

Relational databases have been deeply used by the first digital poetic repertoires combining an ER (Entity-Relationship) model, together with the data model based on records for the logical implementation (Elmasri and Navathe, 2011, 27-ss). The problem of representing ER composition model is that the result shown is data centered, but it is not enough to mark textual items that need to be analyzed from a metrical point of view.

There are other projects based on XML solutions, as TEI has a specific module for poetry analysis, "Verse", with a rich set of tags to describe metrical schemas, rhymes, accentual structure and syllabic varieties. However, this model is not widely used by the different projects, and the lack of philological unified criteria makes it difficult to translate literary schemas into XML tags, making researchers create new tags or express nuances with customized attributes for each project.

The key for interoperability both in philological and technological fields is a common reference system, for which semantic web technologies are a powerful solution. Building a linked data model by adding a semantic layer of metadata to the existing databases does not alter their internal structure. This solution requires, however, to assume unified criteria on the philological model that serves as a reference.

Although semantic web technologies have had success in archives libraries and museums (group known as LODLAM http://lodlam.net/), its application to poetic corpora is very different, as there are only a couple of studies dealing with some of the above mentioned aspects (Bootz and Szoniecky, 2008 and Zöllner-Weber, 2009), but there is not a standard conceptual model of ontology referred to metrics and poetry. The closest works related to this topic are probably the conceptual model of CIDOC, the controlled vocabularies of English Broadside Ballad Project http://ebba.english.ucsb.edu/ and the linked data relations offered by the Library of Congress (http://id.loc.gov/), which do not offer enough information on metrics vocabulary. There are also interesting computational approaches which use automated linguistic analysis or text mining, based on the morphological and phonetic structure of each language. Results have been impressive, as one of the greatest advantages is the speed of the analysis of big amounts of text (Gervás, 2000). Nevertheless, the integration of these technologies with the previous models described is not easy, and solutions are often customized for the variation of natural language used, most times standard English.

Our approach

What we propose in this paper is not a new method for analyzing poetry, but an abstract model based on a working methodology supported by a double standardization system, both at philological and technological levels. In relation to this aspect, it must be highlighted that it needs to be carried out by an interdisciplinary and coordinated team, which requires a careful design of data architecture in different levels. Our proposal aims to set up a procedure to combine philological criteria to map vocabularies and concepts which might have common means and properties in the different traditions and to insert them into an abstract framework in which each of these elements can fit as individuals of an ontology which gathers the main poetics concepts shared by most traditions. We have worked on some first approaches in this sense, building our first ontology prototype, based on our ReMetCa Spanish project: www.purl.org/net/remetca (González-Blanco, and Del Rio, et al., 2014), populated with a controlled vocabulary in SKOS, that can be found in http://vocabularios.caicyt.gov.ar/pmc/. These preliminary results, which served as prototype and a basis to build the current model, have been applied to different poetry projects, such as the edition of *Cancionero de Baena*: http://sade.textgrid.de/exist/apps/SADE/Dialogo_Medieval/index.html or the description of poetic collections in http://www.poetriae.linhd.es/, but they need to be expanded and improved to analyze other poetic systems.

This paper has been developed thanks to the research projects funded by MINECO and led by Elena González-Blanco: Acción Europa Investiga EUIN2013-50630: Repertorio Digital de Poesía Europea (DIREPO) and FFI2014-57961-R. Laboratorio de Innovación en Humanidades Digitales: Edición Digital, Datos Enlazados y Entorno Virtual de Investigación para el trabajo en humanidades, and the Starting Grant ERC-2015-STG-679528 POSTDATA.

Bibliography

Bělič, O. (2000). *Verso español y verso europeo: introducción a la teoría del verso español en el contexto europeo*, Santafé de Bogotá, Instituto Caro y Cuervo.

Bello, A. (1955). *Estudios filológicos. Principios de la ortología y métrica de la lengua castellana y otros escritos en Obras completas* de Andrés Bello, Caracas, Ministerio de Educación. (1st ed. 1835)

Bootz, P. and Szoniecky, S. (2008). Towards an ontology of the field of digital poetry, *Paper presented at Electronic Literature in Europe*. http://elmcip.net/node/415. Accessed: 30/10/2015.

Burnard, L. and Bauman, S. (Eds.) TEI P5: Guidelines for Electronic Text Encoding and Interchange. Ver. 2.5.0. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ Accessed 30-10-2015.

Gerber, N. (2013). Stress-Based Metrics Revisited: A Comparative Exercise in Scansion Systems and their Implications for Iambic Pentameter. *Thinking Verse III*, pp. 131-68.

González-Blanco, E., G. Del Rio, C. Martínez, M. Martos (2014). When TEI Verse becomes linked data: using TEI tags as a model to build a linked poetry system. *Abstract from the paper presented at the TEI Conference 2014*, http://tei.northwestern.edu/files/2014/10/gonzalez-blanco-20wnj3q.pdf

González-Blanco, E. and L. Seláf (2014). Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires. In Soriano, L. et al. (Eds.), *Humanities on the web: the medieval world*. Oxford, Peter Lang, pp. 321-32.

González-Blanco, E. and J. L. Rodríguez (2013). ReMetCa: a TEI based digital repertory on Medieval Spanish poetry, at The Linked TEI: Text Encoding in the Web, *Book of Abstracts - electronic edition*. Edited by F. Ciotti and A. Ciula, DIGILAB Sapienza University and TEI Consortium, Rome 2013, pp. 178-85. http://digilab2.let.uniroma1.it/teiconf2013/abstracts/. Accessed 30-10-2015.

Navarro Tomás, T. (1991). *Métrica española*, Barcelona, Labor. (1st ed., 1956).

Zöllner-Weber, A. (2009). Ontologies and Logic Reasoning as Tools in Humanities? *Digital Humanities Quarterly*, **3**(4). http://www.digitalhumanities.org/dhq/vol/3/4/000068/000068.html Accessed: 30/10/2015.

Translating Electronic Literature. Multicultural, Multilingual and Cross-Platform Encounters

Monika Górska-Olesińska
ikagorska@gmail.com
University of Opole, Poland

Mariusz Pisarski
mariusz.pisarski@techsty.art.pl
University of Warsaw, Poland

Sea and Spar Between by Stephanie Strickland and Nick Montfort is a poetry generator that repurposes, combines and re-mixes words and phrases from Emily Dickinson's poems with those taken from Herman Melville's *Moby Dick* into an (almost) endless sea of stanzas (the program produces about 225 trillion stanzas arranged on a toroidal surface). In his keynote speech to the U.S. Library of Congress, Stuart Moulthrop defined *Sea and Spar Between* as both "an immensely long-form computational poem" and "a remarkably compact poem generator." Transferring such a complex e-literary work into a different language and cultural context raises vital questions regarding the very nature of translation and adaptation in the digital age.

The first Polish version of the program, presented by us in 2013 at Electronic Literature Organization Conference in Paris and published online in 2014 in Techsty (both as a program and as a glossed code with Montfort's and Strickland's comments), greatly multiplied the distributive authorship of the work as a whole, revealing new culturally and linguistically determined aspects of code, grammar and style, and adding a complex layer of interdependencies.

See and Spar Between poses a translational challenge which in some languages might seem impossible to accomplish. Polish, our target language, imposed some serious constraints: one-syllable words became disyllabic or multisyllabic; kennings taken from Melville's work required a different morphological, lexical and grammatical arrangement; and most of the generative rhetoric of the original (like anaphors) had to take into consideration the grammatical gender of Polish words. As a result, the javascript code, instructions that accompany the javascript file, and arrays of words that the poetry generator draws from, needed to be expanded and rewritten. Moreover, at several crucial points of this rule-driven work, the nature of Polish language forced us to modify the code.

In 2015, we started work on porting the *Sea and Spar Between* generator from its javascript+html5 web browser context into XBOX 360 Kinect motion-sensing environment, where the input is controlled by user gestures. This adaptation, which will be available both in English and Polish will be semantically oriented. We aim to use Kinect affordances to program various haptic gestures recognized by the sensor in a way that ensures coherence between interactive gestures and the content of the poem, between the user's haptic activity and their cognitive processes. Moving from one platform to another, from the Web to Kinect, will involve translating mouse and keyboard gestures (point-and-click, numerical input) into a gesture vocabulary of sensor technology. As an adaptation, our work aims at incorporating the dominant themes of Strickland and Montfort's work into user movements during the multidimensional navigation of the work. If the generated stanzas are compared to fish in the ocean, the screen to an infinite canvas and the reader's navigation to a sea voyage, then Kinect port transforms these very metaphors from subject of "translation" into a finite set of gestures that are in sync with the work's semantics and the authorial intentions behind the generator. For this reason alone, the port to Kinect cannot be a translation of a more radical type, in which Dickinson and Melville are replaced with any other poets from any other language.

Once again, the question arises of what we translate/adapt/port when we translate a digital work of art for digitally enhanced venues and for an audience of "digital natives."

In the course of translation of *Sea and Spar Between* and its adaptation for different platforms, the process of negotiation between the source language and the target language involves the factors unseen in traditional translation. Strickland and Montfort read Dickinson and Melville and parse their readings into a computer program, which as a translation, or port, from Python to javascript, is already a derivative. This collision of cultures, languages and tools is amplified when transposed into a different language. The transposition involves the original authors of *Sea and Spar Between*, the original translators of Dickinson and Melville into Polish, and us, turning the process into a multilayered translational challenge, something we propose to call a distributed translation. The forthcoming port to Kinect makes these issues even more challenging and exciting.

Bibliography

Montfort, N., Strickland, S. (2013). Cut to fit the toolspoon course. *Digital Humanities Quarterly*, 7(1).

Montfort, N., Strickland, S. (2014). Spars of Language Find at Sea. *Formules*, 18.

The journal *al-Muqtabas* between *Shamela.ws*, HathiTrust, and GitHub: producing open, collaborative, and fully-referencable digital editions of early Arabic periodicals – with almost no funds

Till Grallert
till.grallert@orient-institut.org
Orient-Institut Beirut, Lebanon

The problems at hand

In the context of the current onslaught cultural artefacts in the Middle East face from the iconoclasts of the Islamic State, from the institutional neglect of states and elites, and from poverty and war, digital preservation efforts promise some relief as well as potential counter narratives. They might also be the only resolve for future education and rebuilding efforts once the wars in Syria, Iraq or Yemen come to an end; and while the digitisation of Archaelogical artefacts has recently received some attention from well-funded international and national organisations, particularly vulnerable collections of texts in libraries, archives, and private homes are destroyed without the world having known about their existence in the first place.[1]

Early Arabic periodicals, such as Butrus al-Bustānī's *al-Jinān* (Beirut, 1876–86), Yaʿqūb Ṣarrūf, Fāris Nimr, and Shāhīn Makāriyūs' *al-Muqtaṭaf* (Beirut and Cairo, 1876–1952), Muḥammad Kurd ʿAlī's *al-Muqtabas* (Cairo and Damascus, 1906–18/19) or Rashīd Riḍā's *al-Manār* (Cairo, 1898–1941) are at the core of the Arabic renaissance (*al-nahḍa*), Arab nationalism, and the Islamic reform movement. These better known and—at the time—widely popular journals do not face the ultimate danger of their last copy being destroyed. Yet, copies are distributed throughout libraries and institutions worldwide. This makes it almost impossible to trace discourses across journals and with the demolition and closure of libraries in the Middle East, they are increasingly accessible to the affluent Western researcher only.[2]

Digitisation seemingly offers an "easy" remedy to the problem of access and some large-scale scanning projects, such as Hathitrust,[3] the British Library's "Endangered Archives Programme" (EAP), MenaDoc or Institut du Monde Arabe produced digital facsimiles of numerous Arabic periodicals. But due to the state of Arabic OCR and the particular difficulties of low-quality fonts, inks, and paper employed at the turn of the twentieth century, these texts can only reliably be digitised by human transcription (c.f. Märgner and El Abed, 2012).[4] Funds for transcribing the tens to hundreds of thousands of pages of an average mundane periodical are simply not available, despite of their cultural significance and unlike for valuable manuscripts and high-brow literature. Consequently, we still have not a single digital scholarly edition of any of these journals.

On the other hand, gray online-libraries of Arabic literature, namely *al-Maktaba al-Shāmila*, *Mishkāt*, *Ṣayd al-Fawāʾid* or *al-Waraq*, provide access to a vast body of, mostly classical, Arabic texts including transcriptions of unknown provenance, editorial principals, and quality for some of the mentioned periodicals. In addition, these gray "editions" lack information linking the digital representation to material originals, namely bibliographic meta-data and page breaks, which makes them almost impossible to employ for scholarly research.

Our proposed solution

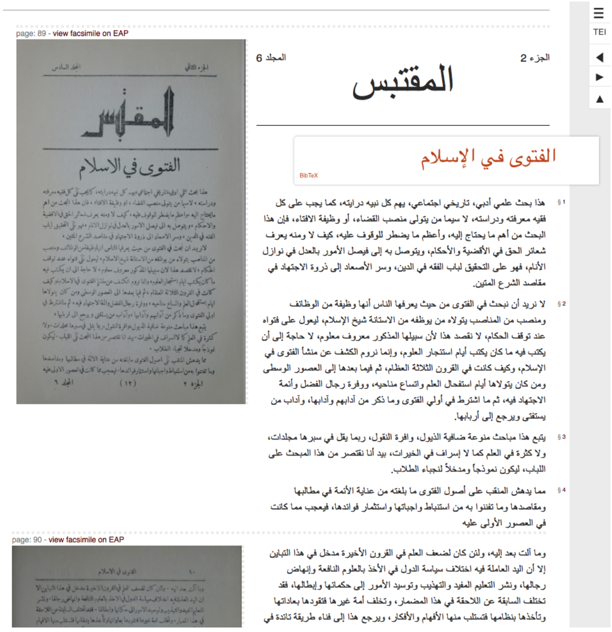
With the GitHub-hosted TEI edition of *Majallat al-Muqtabas*[5] we want to show that through re-purposing well-established open software and by bridging the gap between immensely popular, but non-academic (and, at least under US copyright laws, occasionally illegal) online libraries of volunteers and academic scanning efforts as well as editorial expertise, one can produce scholarly editions that offer solutions for most of the above-mentioned problems—including the absence of expensive infrastructure: We use digital texts from *shamela.ws*, transform them into TEI XML, add light structural mark-up for articles, sections, authors, and bibliographic metadata, and link each page to facsimiles provided through EAP and HathiTrust; the latter step, in the process of which we also make first corrections to the transcription, though trivial, is the most labour-intensive, given that page breaks are commonly ignored by *shamela*.ws's anonymous transcribers. The digital edition (TEI, markdown, and a web-display) is then hosted as a GitHub repository with a CC BY-SA 4.0 licence for reading, contribution, and re-use.[6]

We argue that by linking facsimiles to the digital text, every reader can validate the quality of the transcription against the original we can remove the greatest limitation of crowd-sourced or gray transcriptions and the main source of disciplinary contempt among historians and scholars of the Middle East. Improvements of the transcription and mark-up can be crowd-sourced with clear attribution of authorship and version control using .git and GitHub's core functionality. Such an approach as proposed by Christian Wittern (2013) has recently seen a number of concurrent practical implementations such as project GITenberg led by Seth Woodworth, Jonathan Reeve's Git-lit, and others.

In addition to the TEI XML files we provide structured bibliographic metadata for every article in *al-Muqtabas* (currently as BibTeX). The TEI edition will be referencable down to the word level for scholarly citations, annotation layers, as well as web-applications through a documented and persistent URI scheme.

In order to contribute to the improvement of Arabic OCR algorithms, we will provide corrected transcriptions of the facsimile pages as ground truth to interested research projects starting with transkribus.eu.

To ease access for human readers (the main projected audience of our edition) and the correction process, we also provide a basic web-display that adheres to the principles of GO::DH's Minimal Computing Working group. This web-display is implemented through an adaptation of the TEI Boilerplate XSLT stylesheets to the needs of Arabic texts and the parallel display of facsimiles and the transcription. Based solely on XSLT 1 and CSS, it runs in most internet browsers and can be downloaded, distributed and run locally without any internet connection—an absolute necessity for societies outside the global North.



page: 89 - view facsimile on EAP

المجلد 6 الجزء 2

المقتبس

الفتوى في الإسلام

BibTeX

§1 هذا بحث علمي أدبي، تاريخي اجتماعي، يهم كل نبيه درايته، كما يجب على كل
فقيه معرفته ودراسته، لا سيما من يتولى منصب القضاء، أو وظيفة الافتاء، فإن هذا
البحث من أهم ما يحتاج إليه، وأعظم ما يضطر للوقوف عليه، كيف لا ومنه يعرف
شعائر الحق في الأقضية والأحكام، ويتوصل به إلى فيصل الأمور بالعدل في نوازل
الأنام، فهو على التحقيق لباب الفقه في الدين، وسر الأصعاد إلى ذروة الاجتهاد في
مقاصد الشرع المتين.

§2 لا نريد أن نبحث في الفتوى من حيث يعرفها الناس أنها وظيفة من الوظائف
ومنصب من المناصب يتولاه من يوظفه من الاستانة شيخ الإسلام، ليعول على فتواه
عند توقف الحكام، لا نقصد هذا لأن سبيلها المذكور معروف معلوم، لا حاجة إلى أن
يكتب فيه ما كان يكتب أيام استنجار العلوم، وإنما نروم الكشف عن منشأ الفتوى في
الإسلام، وكيف كانت في القرون الثلاثة العظام، ثم فيما بعدها إلى العصور الوسطى
ومن كان يتولاها أيام استفحال العلم واتساع مناحيه، ووفرة رجال الفضل وأئمة
الاجتهاد فيه، ثم ما اشترط في أولي الفتوى وما ذكر من آدابهم وآدابها، وآداب من
يستفتي ويرجع إلى أربابها.

§3 يتبع هذا مباحث منوعة ضافية الذيول، وافرة النقول، ربما يقل في سبرها مجلدات،
ولا كثرة في العلم كما لا إسراف في الخيرات، بيد أنا نقتصر من هذا المبحث على
اللباب، ليكون نموذجاً ومدخلاً لنجباء الطلاب.

§4 مما يدهش المنقب على أصول الفتوى ما بلغته من عناية الأئمة في مطالبها
ومقاصدها وما تفننوا به من استنباط واجباتها واستثمار فوائدها، فيعجب مما كانت
في العصور الأولى عليه

page: 90 - view facsimile on EAP

وما آلت بعد إليه، ولئن كان لضعف العلم في القرون الأخيرة مدخل في هذا التباين
إلا أن اليد العاملة فيه اختلاف سياسة الدول في الأخذ بالعلوم النافعة وإنهاض
رجالها، ونشر التعليم المفيد والتهذيب وتوسيد الأمور إلى حكمائها وإبطالها، فقد
تختلف السابقة عن اللاحقة في هذا المضمار، وتخلف أمة غيرها فتقودها بعاداتها
وتأخذها بنظامها فتستلب منها الأفهام والأفكار، ويرجع هذا إلى فناء طريقة تائدة في

Figure 1: The web-display of *Digital Muqtabas* based on TEI Boilerplate.

Finally, by sharing all our code, we hope to facilitate similar projects and digital editions of further periodicals. For this purpose, we successfully tested adapting the code to ʿAbd al-Qādir al-Iskandarānī's monthly journal *al-Ḥaqāʾiq* (1910–12, Damascus)[7] in February 2016.

Conclusion

The paper will discuss the challenges cultural artefacts, and particularly texts, face in the Middle East. We will propose a solution to some of these problems based on the principles of openness, simplicity, and adherence to scholarly and technical standards. Applying these principles, our edition of *Majallat al-Muqtabas* improves already existing digital artefacts and makes them accessible for reading and re-use to the scholarly community as well as the general public. Finally, we will discuss the particular challenges and experiences of this still very young project (since October 2015).

Bibliography

Commins, D. (1990). *Islamic Reform: Politics and Social Change in Late Ottoman Syria*. Oxford: Oxford University Press.

Glaß, D. (2004). *Der Muqtaṭaf Und Seine Öffentlichkeit. Aufklärung, Räsonnement Und Meinungsstreit in Der Frühen Arabischen Zeitschriftenkommunikation*. Würzburg: Ergon Verlag.

Grallert, T. (2013). The puzzle continues: al-Muqtaṭaf was printed in two different and unmarked editions, *Sitzextase,* http://tillgrallert.github.io/blog/2013/08/19/the-puzzle-continues/ (accessed 6 February 2016).

Grallert, T. (2014). The puzzle continues II: in addition to al-kabīr and al-ṣaghīr, al-Muqtaṭaf published slightly different editions in Beirut and Kairo, *Sitzextase,* http://tillgrallert.github.io/blog/2014/01/19/the-puzzle-continues-2/ (accessed 6 February 2016).

Märgner, V. and El Abed, H. (Eds) (2012). *Guide to OCR for Arabic Scripts*. London: Springer http://link.springer.com/book/10.1007/978-1-4471-4072-6.

Seikaly, S. (1981). Damascene Intellectual Life in the Opening Years of the 20th Century: Muhammad Kurd ʿAli and Al-Muqtabas. In Buheiry, M. R. (ed), *Intellectual Life In The Arab East, 1890-1939*. Beirut: American University Of Beirut, pp. 125–53.

Wittern, C. (2013). Beyond TEI: Returning the Text to the Reader. *Journal of the Text Encoding Initiative*, 4: Selected Papers from the 2011 TEI Conference. http://jtei.revues.org/691.

Notes

[1] For a good example of crowd-sourced conservation efforts targetted at the Armenian communities of the Ottoman Empire see the Houshamadyan project, which was established by Elke Hartmann and Vahé Tachjian in Berlin in 2010 and launched an "Open Digital Archive" in 2015. Other digitisation projects worth mentioning are the Yemen Manuscript Digitisation Project (University of Oregon, Princeton University, Freie Universität Berlin) and the recent "Million Image Database Project" of the Digital Archaeology Institute (UNESCO, University of Oxford, government of the UAE) that aims at delivering 5000 3D cameras to the MENA region in spring 2016.

[2] In many instances libraries hold incomplete collections and only single copies. This, for instance, has caused even scholars working on individual journals to miss the fact that the very journal they were concerned with appeared in at least two different editions (e.g. (Glaß, 2004) see (Grallert, 2013; Grallert, 2014)).

[3] It must be noted that the US-based HathiTrust does not provide public or open access to its collections even to material deemed in the public domain under extremely strict US copyright laws to users outside the USA. Citing the absence of editors able to read many of the languages written in non-Latin scripts, HathiTrust tends to be extra cautious with the material of interest to us and restricts access by default to US-IPs. These restrictions can be lifted on a case-by-case basis, which requires at least an English email conversation and prevents access to the collection for many of the communities who produced these cultural artefacts; see https://www.hathitrust.org/access_use for the access policies.

[4] For the abominable state of Arabic OCR even for well-funded corporations and projects, try searching inside Arabic works on Google Books or HathiTrust.

[5] For a history of Muḥammad Kurd ʿAlī's journal *al-Muqtabas* (The Digest) see (Seikaly, 1981) and the readme.md of the project's GitHub repository.

[6] The text of *al-Muqtabas* itself is in the public domain even under the most restrictive definitions (i.e. in the USA); the anonymous original transcribers at *shamela.ws* do not claim copyright; and we only link to publicly accessible facsimile's without copying or downloading them.

[7] On the history of *al-Ḥaqāʾiq* and some of its quarrels with *al-Muqtabas* see (Commins, 1990:118–22).

Archives Distant Reading: Mapping the Activity of the League of Nations' Intellectual Cooperation

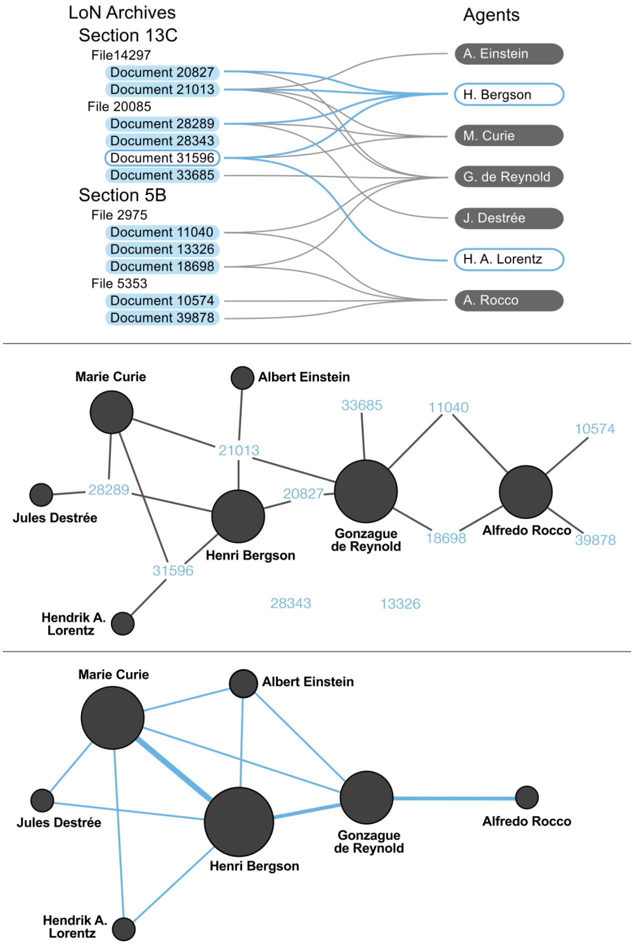
Martin Grandjean
martin.grandjean@unil.ch
University of Lausanne, Switzerland

Introduction

Founded in 1922 by the League of Nations upon observation that the pacification of Europe may benefit from a better collaboration between scientific elites, the *International Committee on Intellectual Cooperation* (ICIC) is responsible for coordinating the restructuration of knowledge circulation. Bringing together leading researchers at the height of their career, such as Albert Einstein, Marie Curie and George Hale, chaired by Henri Bergson, the Committee weaves a complex network between transnational scientific institutions and societies, congresses and individuals (Pernet, 2014).

This paper proposes an analysis of the work and functioning of the organization between 1919 and 1927 by setting up a database containing metadata of thousands of documents contained by the ICIC funds (United Nations Archives, Geneva). Visualized as a network of 3.200 people (tens of thousands of relationships), this work provides a new understanding of the internal organization of the Intellectual Cooperation, as well as completely new insights about its relations with the rest of the scientific and diplomatic world. In particular, we will show the necessity to compare the “micro” structure of relationships as mapped by the archive with the “macro” formal structure of the institution. Do the thousands of documents, in a *distant reading* approach (Moretti, 2013), confirm the internal organization of the League of Nations or do they show individuals/communities that bypass the official hierarchy?

As an opening to an epistemological debate, this research questions the relationship between the researcher, the database and its sources: are the metadata of an archive corpus usable information, regardless of their unique qualitative content? More technically, it also addresses the issue of data visualization and modeling in the historical sciences.



Figures 1-3. From a relational database to a network. The relational database links documents with their agents (Fig. 1), a relation that is then mapped as a 2-mode network (Fig. 2). By projecting the documents on the agents, the 1-mode network of co-occurrencies is fully exploitable (Fig. 3).

Sources

Initially launched in 1922 as a consultative commission, the ICIC quickly mobilized the major part of the *International Bureaux* section secretariat (upon its stabilization a few years later as a permanent commission), resulting in the production of very vast archives. In the period covered by our study, 1919-1927, which can be qualified as the start-up years of the dynamics of intellectual cooperation in Europe and the World, the funds contain 27.000 documents, mostly internal and external correspondence, about the main missions of the commission: university relations, bibliography coordination, educational matters and various enquiries.

Methodology

We are particularly interested in individuals who are personally concerned by the documents. Firstly, we indexed all the documents by creating a relational database (fig.1) of all "agents" (senders and receivers). In order to analyze the co-occurrences of agents in the same document, the database, displayed as a 2-mode network (fig.2), is projected onto a 1-mode network (fig.3). Each of the 3.200 agents are connected to its co-occurring by an edge whose weight reflects the intensity of the relationship.

Data analysis and visualization

This paper presents the final result of years of manual indexing (intermediary results have already been presented as case-studies in Grandjean, 2015 and Grandjean, 2014). The complete graph (fig.4) displays 38.600 co-occurrences between 3.200 agents of the complete set of documents from 1919 to 1927. The size and color of the nodes are proportional to the number of appearances of the individuals in the index. The size and color of the edges are proportional to the number of co-occurrences of the two people they bind.

Beyond an apparently low visual intelligibility, due to the amount of information and the dataset complexity, such a graph already makes the measurement of its mathematical properties possible (centrality measures, as detailed by Koschützki et al., 2005 or Newman, 2010). Developed as an interactive online visualization, it provides a global view and a more instinctive navigation in the archive directory.

This type of graph is necessary in order to observe what happens at the margins of the institution (and thus also to understand the geography of the object: what is central and what is peripheral). However, as so often in network analysis, the core is so dense that we can not distinguish the edges and therefore makes a more advanced visual analysis impossible.

Here comes the challenge of readability: how to show that elements are not only connected horizontally to other elements by maintaining a macro-structure that does not always correspond to the natural organization of the agents?

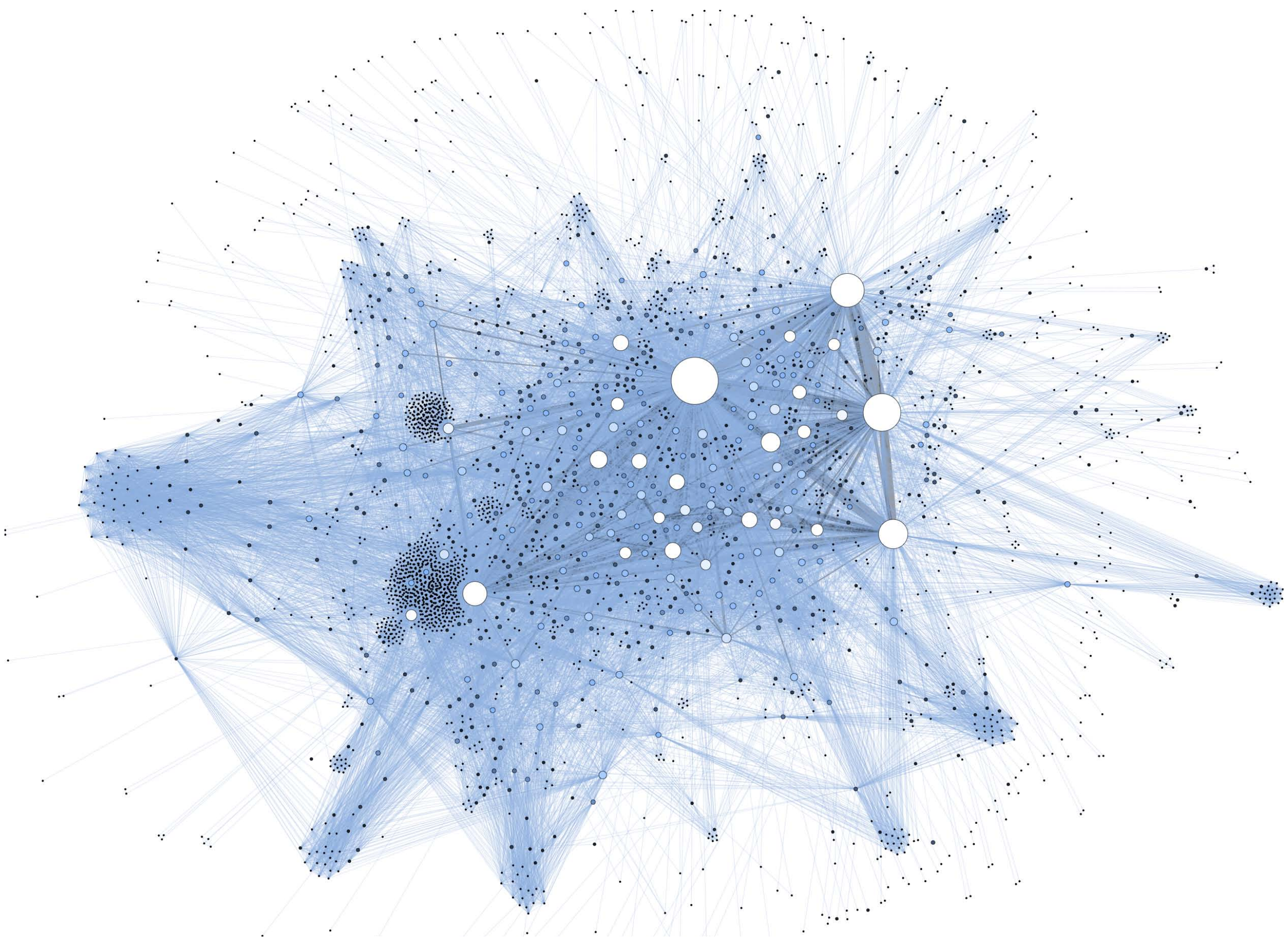


Figure 4. The network of the Intellectual Cooperation. 38.6K co-occurrences of 3.200 agents of the documents in the League of Nations' Intellectual Cooperation archives (1919-1927, 27K documents).

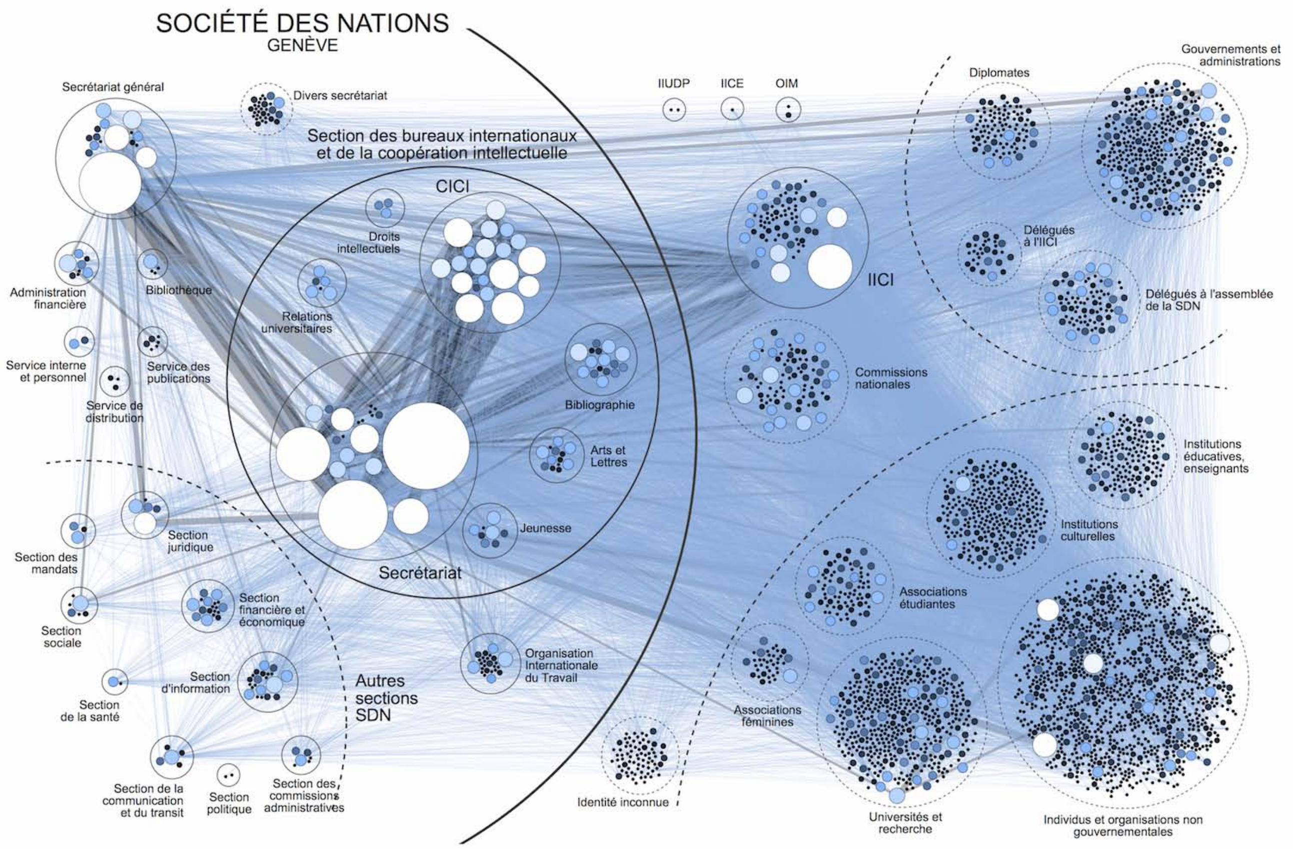


Figure 5. Untangling the network. Same network as Fig.4 but mapped according to the institutional structure of the League of Nations, introducing a 3rd dimension.

As many opportunities to play with scale exist (Brailly and Lazega, 2012), we chose to flatten the institutional organization on the relational structure (fig.5). Hence our research question: do the scientists, diplomats and senior officials - who constitute the network of intellectual cooperation - structure their relationships in coherence with the organization of their institution? Or, are they the ones that determine the links that their institutions maintain together? Forced distribution of nodes under an administrative "geography" make it easier to read the edges between groups. We also note that this is a way to provide a spatial distribution that does not vary over time, and thus allows the study of several successive moments of the network, without losing the mental map. This system is also suitable for superimposing prosopographical information.

Perspectives

It is often at the periphery of the network that the most interesting personal trajectories may be found. As such, this display allows us to discover and highlight the thematic affinities of some of the privileged interlocutors of the ICIC: government delegates, heads of international scientific organizations or partners seeking asylum under the authority of the League of Nations.

The consistency of many internal networks can be evaluated: is the plenary commission a clique (cluster where each and every node is connected to all the others)? Are the expert sub-committees coherent communities? And are these communities created by a well defined group of documents or by a heterogeneous collection of correspondences on various themes?

Pushing further the distant reading, we will also see that the macro-analysis of the institutional level reveals structures that were not clear at the individual level, showing the need for a constant back-and-forth between the scales. This will also be an opportunity to recall that network analysis is a modeling process that does not relieve the researcher of the consultation of the archival documents themselves. From a quantitative approach, we return to the qualitative: the structural organization of a network is definitely a "qualitative, morphological" information (Moretti, 1999; 68) derived from the quantitative compilation of individual relationships.

Bibliography

Brailly, J. and Lazega E. (2012). Diversité des approches de modélisation statistique en analyse de réseaux sociaux multiniveaux, *Mathematics and Social Sciences*, **198**: 5-28.

Grandjean, M. (2014). La connaissance est un réseau, perspectives sur l'organisation archivistique et encyclopédique, *Les Cahiers du Numérique*, **10**(3): 37-54.

Grandjean, M. (2015). Introduction à la visualisation de données : l'analyse de réseau en histoire, *Geschichte und Informatik*, **18**(19): 109-28.

Koschützki, D., et al. (2005). Centrality Indices. In Brandes U. and Erlebach T. *Network Analysis*. Springer, pp. 16-61.

Moretti, F. (1999). *Atlas of the European Novel, 1800-1900*, Verso Books.

Moretti, F. (2013). *Distant reading*, Verso Books.

Newman, M. E. (2010). *Networks: An Introduction*, Oxford University Press.

Pernet, C. (2014).

Twists, Turns and Dead Alleys: The League of Nations and Intellectual Cooperation in Times of War, *Journal of Modern European History*, **12**(3): 342-58.

Scholarly Requirements for Large Scale Text Analysis: A User Needs Assessment by the HathiTrust Research Center

Harriett Elizabeth Green
green19@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Eleanor Frances Dickson
dicksone@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Sayan Bhattacharyya
sayan@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Introduction

The HathiTrust Research Center (HTRC) aims to facilitate large-scale computational text analysis of the contents of the HathiTrust Digital Library (HTDL) through data services and analytical tools. We conducted a study of current and potential users of the HTRC to investigate how scholars integrate text analysis into their research. Our study aims to inform the development of HTRC services and also to generate deeper insights into scholarly research practices with large-scale digitized text corpora.

Background

Studies on the use of digital content by humanities scholars, ranging from humanities cyberinfrastructure (ACLS, 2006) and patterns in scholarly practices (Brockman et al., 2001; Palmer and Neumann, 2002; Green and Courtney, 2015), to discipline-specific studies

(Zorich, 2012; Babeu, 2011; Rutner and Schonfeld, 2011), reveal that scholars acquire and analyze digital content in multi-faceted ways. Several investigations particularly examine scholarly uses of digital tools (Frischer et al., 2006; Toms and O'Brien, 2008; Gibbs and Owens, 2012). Computational text analysis dates from the beginnings of humanities computing (Hindley, 2013), and the resources of the ARTFL Project (Argamon et al., 2009; Horton et al., 2009), MONK (Unsworth, 2011), Wordseer (Muralidharan and Hearst, 2013), Voyant and TaPOR (Rockwell et al., 2010), and Lexos (LeBlanc et al., 2013), among others, inform the current work of the HTRC to provide a secure computational and data environment for researchers to conduct analyses of content from the HathiTrust Digital Library.

Our study builds on an earlier user needs assessment conducted for the HTRC and its Mellon Foundation-funded Workset Creation for Scholarly Analysis project. That earlier study analyzed interviews and focus groups in order to identify capabilities needed in large text corpora to facilitate scholarly research use (Fenlon et al., 2014). These desired capabilities included the ability to create and manipulate collections as reusable datasets and research products, the ability to work at different units of analysis, and access to highly enriched metadata (Green et al., 2014; Fenlon et al., 2014).

Our present study especially builds upon that previous investigation by examining the text analysis research practices of current and potential users of the HTRC.

Research Design

Goals

Our study's primary goals are:

- To analyze current scholarly research practices with textual corpora to identify user requirements for HTRC services;
- to develop illustrative use cases of text analysis research for shaping training curricula; and
- to obtain information for guiding the development of HTRC research services in the University of Illinois Library's Scholarly Commons and similar digital scholarship centers.

While the findings of this study specifically will inform the development of services to meet the needs of HTRC users, it also contributes broader insights into how to develop similar digital resources and research services for computational text analysis.

Methods

We conducted fifteen semi-structured interviews with students, faculty, researchers, administrators, and librarians who pursue work that includes text analysis, or have familiarity with text analysis methods. Some participants were recruited at professional conferences for digital humanities and libraries, while others were active in HTRC user group forums. Several of the interviewees had previously interacted with the HTRC, and most had experience with the HTDL. The participants were from various disciplines — including English, Anthropology, History, and Computer Science —and ranged from newcomers to digital humanities to long-time researchers.

We performed an initial analysis of the interview data through open coding and will continue detailed qualitative analysis using ATLAS.ti. Data was independently coded by the authors to ensure inter-coder reliability. While we are still actively analyzing interview data, we identified several preliminary themes discussed here. These themes include strategies for obtaining and managing data, research workflows and results, collaborations, and teaching.

Analysis and Discussion

Data Acquisition and Management

Several respondents characterized text analysis research as being time-intensive in spite of the speed of computational tools. One interviewee noted, 'It's funny, often people think, "Oh we have it digitized, now it's useful." Scholars realize that you have a lot more work to do after that. And that can often slow projects down terribly.'

The interviewees indicated that gathering, managing, and manipulating text data comprised a considerable portion of their work. An interviewee explained, 'I think the biggest challenge is data, getting good data to work with. I think people underestimate the problems and difficulties in doing that.'

Interviewees also expressed a desire for improved ways to identify and extract the content they needed, especially when navigating large-scale collections to find the volumes, pages, or passages relevant to a research project. As one interviewee remarked, 'Even if you had somehow structured your texts, I would be saying, "What was left out? How do I bring it back in?"'

Research Workflows and Results

Several interviewees described the potential of text analysis to challenge previously held understandings of text, as differences between human and computational readings emerged. One respondent noted, 'There are many cases in which the computer is at least as good—if not better—a reader than humans are. That's very difficult for people to accept... sometimes the computer gets it right and it bears looking at that difference. So we kind of want to get that new ground truth on this kind of work.'

Many researchers highlighted the importance of interpretive work in understanding how the tools interact with the text, and characterized the interactions as dynamic. One respondent observed, 'I yearn for workflows where the scholar could actually set their own tokenization rules.... It would be a way that we could create less language-specific [rules] or control the language specificity of the algorithm. I think that is the real need.' Several respondents highlighted the importance of tools that flexibly fit into various stages of the research process, and also are accessible to users of different skill levels. Interviewees also suggested enhancements specific to the HTRC, which included expanded visualization capabilities, improved generation of statistics about text corpora, and better ability to handle languages other than English.

Research Collaborations

Interviewees repeatedly cited collaboration and research support, both virtual and in-person, as important. Many interviewees worked with digital humanities initiatives, and reported that their local resources ranged from limited technical support to well-resourced research centers. For some interviewees, online support communities— such as Digital Humanities Questions and Answers or Stack Overflow — also were significant.

Interdisciplinary collaborations between departments and across institutions emerged as the most prominent kind of partnership, but interviewees also noted the challenges that such collaborations pose. As one interviewee explained, 'Collaborations between institutions: much more difficult. There's money, there's institutional blockages, and then anything over half a dozen people, it gets complicated very quickly. And so the people dynamics get very complicated.' Some respondents noted that these collaborations affected their research practices and acquisition of research resources.

Interviewees reported that their collaborations with libraries ranged from non-existent to critical partnerships. Many saw the library as a key space because 'the library is actually the one functioning interdisciplinary space on a university campus.' Collaborations with the HTRC and digital repositories for working with data also were important to respondents.

Teaching and Training

Interviewees mentioned their active efforts and intentions to incorporate computational text analysis into their teaching. Some remarked on institutional constraints that make it difficult to incorporate computational tools into curricula. As one respondent explained: 'I once imagined teaching a class in which students learn to script and actually run analyses against data, but I was told, basically, that that class isn't a humanities class anymore—that belongs in computer science.'

Some stated that the courses that they currently teach may not require or allow for the incorporation of computational analysis. Yet others noted that there is only a limited amount of technical or scientific skills that a humanities student could realistically master within a short period of time, with one interviewee noting that 'you can only get people to learn so much about the math; as much as they can learn, they should — at the same time, it's hard.'

Although the demand from students for learning about computational text analysis was, overall, reported to be increasing, some interviewees noted that they are constrained by not only limited resources, but also uncertainty as to how to carry out such activities. One interviewee reported prevailing sentiments that the digital humanities 'doesn't even fit anywhere,' leading to the question of whether 'there should be a whole separate department that's digital humanities,' or to offer training within existing curricula.

Conclusion

The immediate aims of this study are to generate an updated framework of user requirements that will guide the development of the HTRC's educational programming and research support services and also to inform forthcoming Mellon Foundation-funded development of the HTRC Data Capsule. But our preliminary findings also provide insights into scholars' needs as they increasingly incorporate text analysis in research and teaching. These findings also reveal how digital scholarship centers, information professionals, and providers of digitized content can best support scholarship as digital humanities resources evolve.

Acknowledgements

We thank Megan Senseney, Angela Courtney, Nicholae Cline, and Leanne Mobley for their collaboration in this study.

Bibliography

American Council of Learned Societies. (2006). *Our Cultural Commonwealth: The report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences*. New York: American Council of Learned Societies. http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf (accessed 4 March 2016).

Argamon, S., et al. (2009). Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters. *Digital Humanities Quarterly* **3**(2): http://www.digitalhumanities.org/dhq/vol/3/2/000043/000043.html

Babeu, A. (2011). *Rome wasn't digitized in a day: Building a cyberinfrastructure for digital classics.* CLIR Publication, **150**, Washington, DC: Council of Library and Information Resources. http://www.clir.org/pubs/reports/pub150/reports/pub150/pub150.pdf (accessed 4 March 2016).

Brockman, W. S., et al. (2001). *Scholarly work in the humanities and the evolving information environment* CLIR Publication, **104**, Washington, D.C.: Digital Library Federation, Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub104/pub104.pdf (accessed 4 March 2016).

Fenlon, K., et al. (2014). Scholar-built collections: A study of user requirements for research in large-scale digital libraries. *Proceedings of the American Society for Information Science and Technology*, **51**(1): 1–10.

Frischer, B., et al. (2006). *Summit on digital tools for the humanities: Report on summit accomplishments*. Institute for Advanced Technology in the Humanities, University of Virginia. http://www.iath.virginia.edu/dtsummit/SummitText.pdf (accessed 4 March 2016).

Gibbs, F. and Owens, T. (2012). Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly*, **6**(2).http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html (accessed 4 March 2016).

Green, H. and Courtney, A. (2015). Beyond the Scanned Image: A Needs Assessment of Scholarly Users of Digital Collections. *College and Research Libraries*, **76**(5): 690-707.

Green, H. E., et al., (2014). Using Collections and Worksets in Large-Scale Corpora: Preliminary Findings from the Workset Creation for Scholarly Analysis Prototyping Project. *Poster presented at iConference 2014*, Berlin, Germany.

Hindley, M. (2013). The Rise of the Machines: NEH and the Digital Humanities: the early years. *Humanities*, **34**(4).http://www.neh.gov/humanities/2013/julyaugust/feature/the-rise-the-machines (accessed 4 March 2016).

Horton, R., et al. (2009). Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie. *Digital Humanities Quarterly*, **3**(2): http://www.digitalhumanities.org/dhq/vol/3/2/000044/000044.html

LeBlanc, M. D., et al. (2013). Lexomics: Integrating the Research and Teaching Spaces. *Digital Humanities 2013 Conference Abstracts, University of Nebraska–Lincoln, 16-19 July 2013*. Lincoln, NE: Association of Digital Humanities Organizations, pp. 274-76. http://dh2013.unl.edu/abstracts/ab-293.html (accessed 4 March 2016).

Muralidharan, A. and Hearst, M. A. (2013). Supporting Exploratory Text Analysis in Literature Study.*Literary and Linguistic Computing*, **28**(2): 283-95. 10.1093/llc/fqs044.

Palmer, C. L. and Neumann, L. J. (2002). The Information Work of Interdisciplinary Humanities Scholars: Exploration and Translation. *Library Quarterly* **7**(1): 85-117.

Rockwell, G., et al. (2010). Ubiquitous Text Analysis. *Poetess Archive Journal*, **1**(2).https://journals.tdl.org/paj/index.php/paj/article/view/13 (accessed 4 March 2016).

Rutner, J. and Schonfeld, R. (2012). *Supporting the Changing Research Practices of Historians*. New York: Ithaka S+R. http://sr.ithaka.org/?p=22532

Sukovic, S. (2011). E-Texts in Research Projects in the Humanities, A. Woodsworth and W. D. Penniman (eds.), *Advances in Librarianship*. Bingley, UK: Emerald Group Publishing, pp. 131-202.

Toms, E. G. and O'Brien, H. (2008). Understanding the Information and Communication Technology Needs of the E-Humanist. *Journal of Documentation*, **64**(1): 102-30.

Unsworth, J. (2011). Computational Work with Very Large Text Collections: Interoperability, Sustainability, and the TEI. *Journal of the Text Encoding Initiative* **1**, 10.4000/jtei.215 (accessed 4 March 2016).

Zorich, D. (2012). *Transitioning to a Digital World: Art History, Its Research Centers and Digital Scholarship: A Report to the Samuel H. Kress Foundation and the Roy Rosenzweig Center for History and New Media*. New York: Samuel H. Kress Foundation. http://www.kressfoundation.org/research/Default.aspx?id=35379 (accessed 4 March 2016).

New Maps for the Lettered City: a Data Visualization Exploration of 19th Century Salons in Mexico

Silvia Eunice Gutiérrez De la Torre
silviaegt@gmail.com
Würzburg Universität, Germany

Visualizations have a central role in the Digital Humanities. The second most popular author-chosen topic word of DH2015 was visualization (cf. Figure 1 from Weingart, 2015). Yet, when one revisits all accepted abstracts at DH2015 with the keyword "vis(z)ualization", one may notice that not many of these texts indicate which libraries were used for interactivity (only some mention D3), and even fewer had direct links to their websites for testing (many were prototypes).

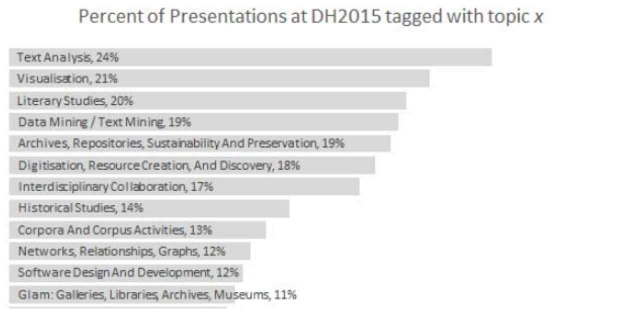


Figure 1. Fragment from graphic with the topical coverage of DH2015 (Weingart, 2015)

One of the more common visualizations are those of relational data. In Katrien Verbert's more thorough survey of interactive visuals in DH2014's presentations, relations-visualizations represent 50% of all prototypes (that is 29 out of 58). The most popular way to represent relationships are uni- or bi-directional graphs (23 out of 29) and only one of them used a matrix to display connections (see figure 2 from Verbert, 2015). To crack open the discussion on the pros and cons of this visualization technique, I will show how I answered some questions of cultural history, more specifically of Latin American Literary History, with a tailored interactive matrix (to this the visualizations visit: http://www.sgutierrez.seewes.de/).

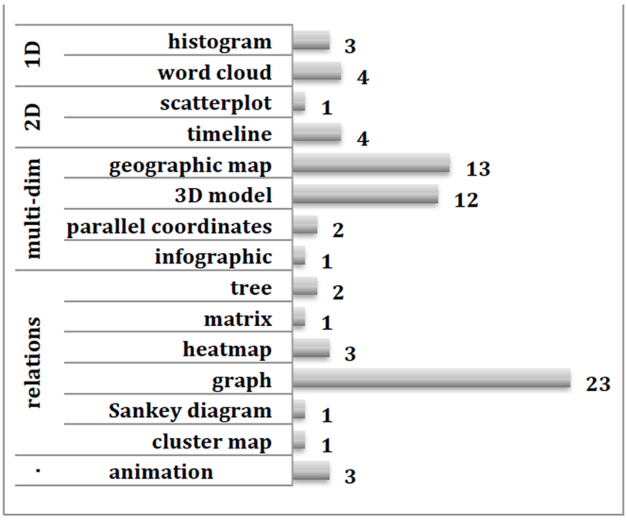


Figure 2. Visualization techniques used by work presented at DH14 (Verbert, 2015)

Theoretical Framework

The history of associations is a goldmine for the intellectual queries of scholars interested in literary and intellectual history. In Latin America, for instance, the appearance of Angel Rama's posthumous book *The lettered city* (1984), lead to a series of studies concerned with the constitution of enlightened groups, especially in new nation-states. As the capital city of one of the most powerful ex-Spanish Colonies, the lettered network in Mexico City makes an interesting case study. However, despite the valuable monographic studies on this subject (most notably Perales Ojeda, 2000 and Sánchez, 1951), which register around 200 active literary societies during the 19th century, no overview on the subject has been possible and not all questions have been resolved. How diverse or homogeneous were these groups? Who were their most recurrent actors? Were certain generations more likely to be part of groups from a certain literary movement? I will propose a way of using visual and interactive displays of literary societies' membership data to answer these three questions.

Before me, others have sought to gain new insights by exploring the possibilities of data modeling to understand modern sociability. The *Berliner Klassik Gesseligkeit Datenbank* (The Societies Database of the Berlin Classical Period, 2013) aimed to understand the cultural bloom of the early 19th century and Stanford's *Salons Project* (2012) was designed to get an understanding of the social composition of the French Enlightenment network. However, to date, there are no online dynamic visualizations of either of these projects.

Methodology

a) data collection

The network's information was obtained by scrapping each associations' entry in the Encyclopedia of Mexican Literature (ELEM). Since ELEM is the most complete source of biographical data for 19th century Mexican writers, it is very unlikely that information about these writers can be found elsewhere; thus, I only considered members with an entry in this source. This procedure leaves out many characters, but it is at least representative of the known characters of the lettered city. It contains information of 51 associations (founded between 1808 and 1894) and of 195 members born between 1781 and 1870.

b) data model

The database derived from these two nodes (members and associations) was modelled to answer my research questions, but its metadata is designed to be reusable: members were assigned standard identifiers using Jeff Chiu's VIAF reconciliation service for OpenRefine (Chiu, 2015), and neutral aspects about these nodes—such as birth and death dates or founding and closing dates—were included. In addition to these neutral aspects, I added two categories that scholars have used to cluster literary characters and societies, namely, generations and literary movements.

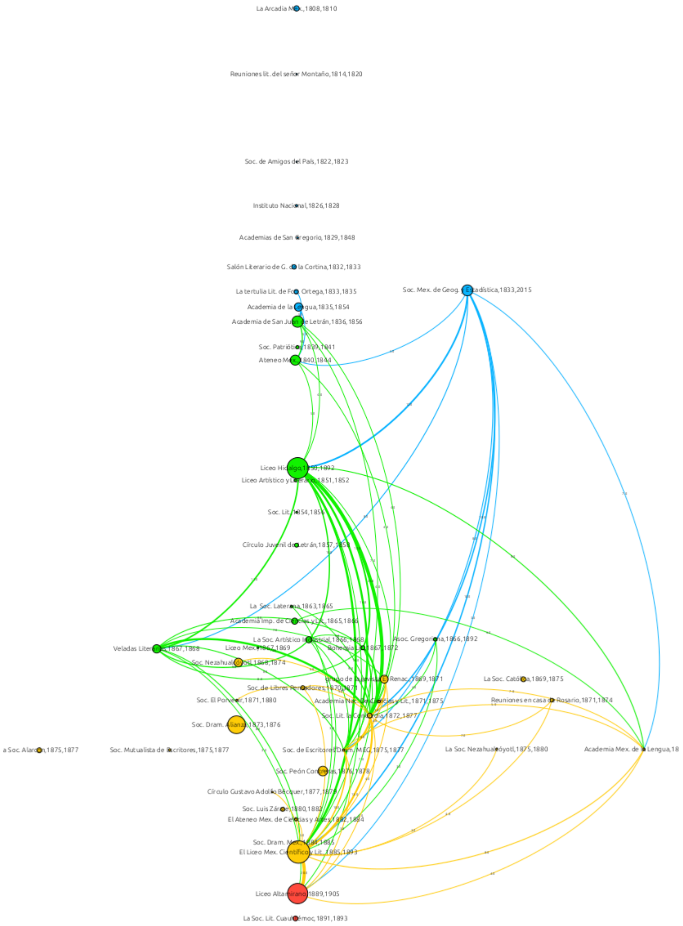

La Arcadia Mex.,1808,1810
Reuniones lit. del señor Montaño,1814,1820
Soc. de Amigos del País,1822,1823
Instituto Nacional,1826,1828
Academias de San Gregorio,1829,1848
Salón Literario de G. de la Cortina,1832,1833
La tertulia Lit. de Fco. Ortega,1833,1835
Soc. Mex. de Geog. y Estadística,1833,2015
Academia de la Lengua,1835,1854
Academia de San Juan de Letrán,1836,1856
Soc. Patriótica,1839,1841
Ateneo Mex.,1840,1844
Liceo Hidalgo,1850,1892
Liceo Artístico y Literario,1851,1852
Soc. Lit.,1854,1856
Círculo Juvenil de Letrán,1857,1858
La Soc. Laterana,1863,1865
Academia Imp. de Ciencias y Lit.,1865,1866
Veladas Literarias,1867,1868
La Soc. Artística Industrial,1866,1868
Asoc. Gregoriana,1866,1892
Liceo Mex.,1867,1869
Bohemia Lit.,1867,1872
Soc. Nezahualcóyotl,1868,1874
grupo de la revista El Renac.,1869,1871
La Soc. Católica,1869,1875
Soc. de Libres Pensadores,1870,1871
Soc. El Porvenir,1871,1880
Academia Nac. de Ciencias y Lit.,1871,1875
Reuniones en casa de Rosario,1871,1874
Soc. Lit. la Concordia,1872,1877
Soc. Dram. Alianza,1873,1876
a Soc. Alarcón,1875,1877
Soc. Mutualista de Escritores,1875,1877
Soc. de Escritores Dram. MEG,1875,1877
La Soc. Nezahualcóyotl,1875,1880
Academia Mex. de la Lengua,18
Soc. Peón Contreras,1876,1878
Círculo Gustavo Adolfo Bécquer,1877,1879
Soc. Luis Zárate,1880,1882
El Ateneo Mex. de Ciencias y Artes,1882,1884
Soc. Dram. Mex.,1884,1885
El Liceo Mex. Científico y Lit.,1885,1893
Liceo Altamirano,1889,1905
La Soc. Lit. Cuauhtémoc,1891,1893

Figure 3. Network visualization where nodes are 19th century Mexican societies and edges represent the number of common members between them

c) visualization

My first attempt was to follow the most common visualization for networks in the digital humanities, the Gephi-spaguetti (see figure 3). I did everything I could to enhance readability. I set the societies-nodes' size according to the number of connections they had with other associations and the thickness of the edges to vary depending on how many common members two groups had. Even more, in order to get a chronologically-ordered layout I used Spekkink's useful plugin, the Event Graph Layout (Spekkink, 2014). From this display, I was able to confirm kinship-relationships between societies. That is, that although persistence was never their *forte*, when one looks at the number of members that went from an extinct society to the next new one, one gets the impression that despite the ephemeral nature of these groups, there was still a type of continuity among them. Yet, even when I created an interactive graph with Sigma.js it was very hard to read the quantitative differences between my nodes' connections. On the one side, I was interested in creating a visual display that allowed interactivity, providing end-users with both additional information for each data-point and the possibility to select specific ranges of the network. On the other, I wanted to control the order of my data and the quantities' color-coding for readability. The solution was provided by a Python-library, Bokeh.

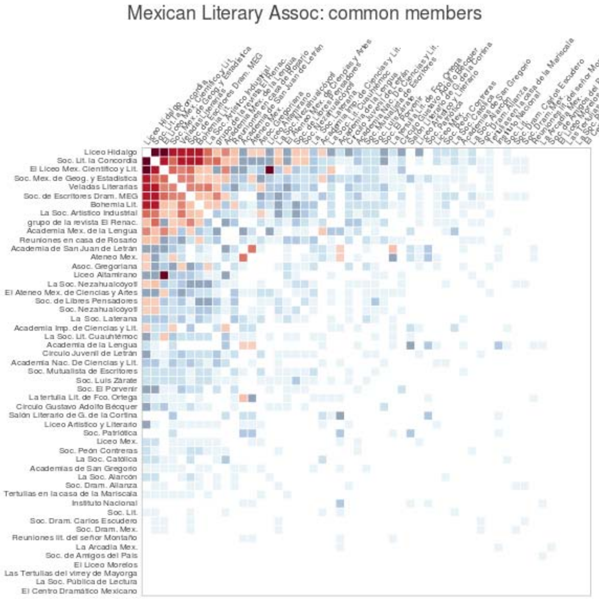


Figure 4. Co-occurrence matrix of literary societies ordered by the sum of common members with other associations

Results

The first visual I created was a co-occurrence matrix where each literary association was compared against all others. This display allowed me to understand how many members each pair of associations had in common. In order to enhance the identification of meaningful co-occurrences, I followed the principles of sequential color schemes –where low data values are represented by light colors and high values by dark ones (Wyssen, 2014) – and I assigned different colors and alphas according to the quantities' distribution of my data: associations' pairs above or equal to five common members were coded in red, and below five, in blue (see Figure 4). Additionally, I set different and consecutive alpha values to each glyph according to their exact value (intersections of less density had lower alphas). This display was helpful to address the question on the diversity or homogeneity of literary societies: with this tailored visual I was able to identify the homogenous hub of ten literary associations around the *Liceo Hidalgo* that had a considerable amount of common members, suggesting that although they had different approaches they were nonetheless constituted by recurring members (cf. Figure 5).

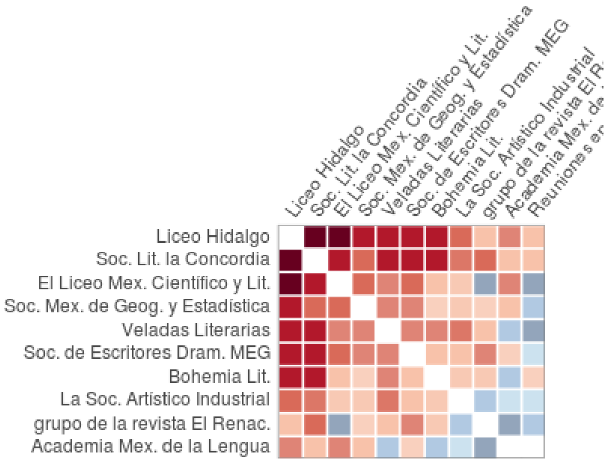


Figure 5. Selection of societies with the highest common-members' density

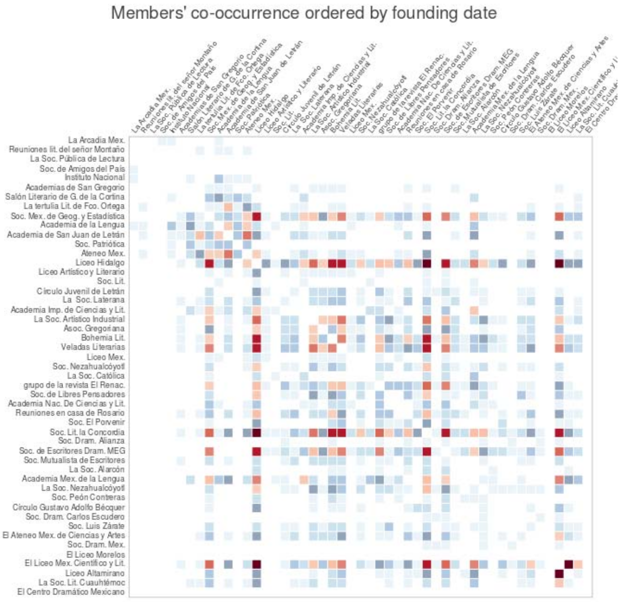


Figure 6. Associations' co-occurrence matrix by founding date

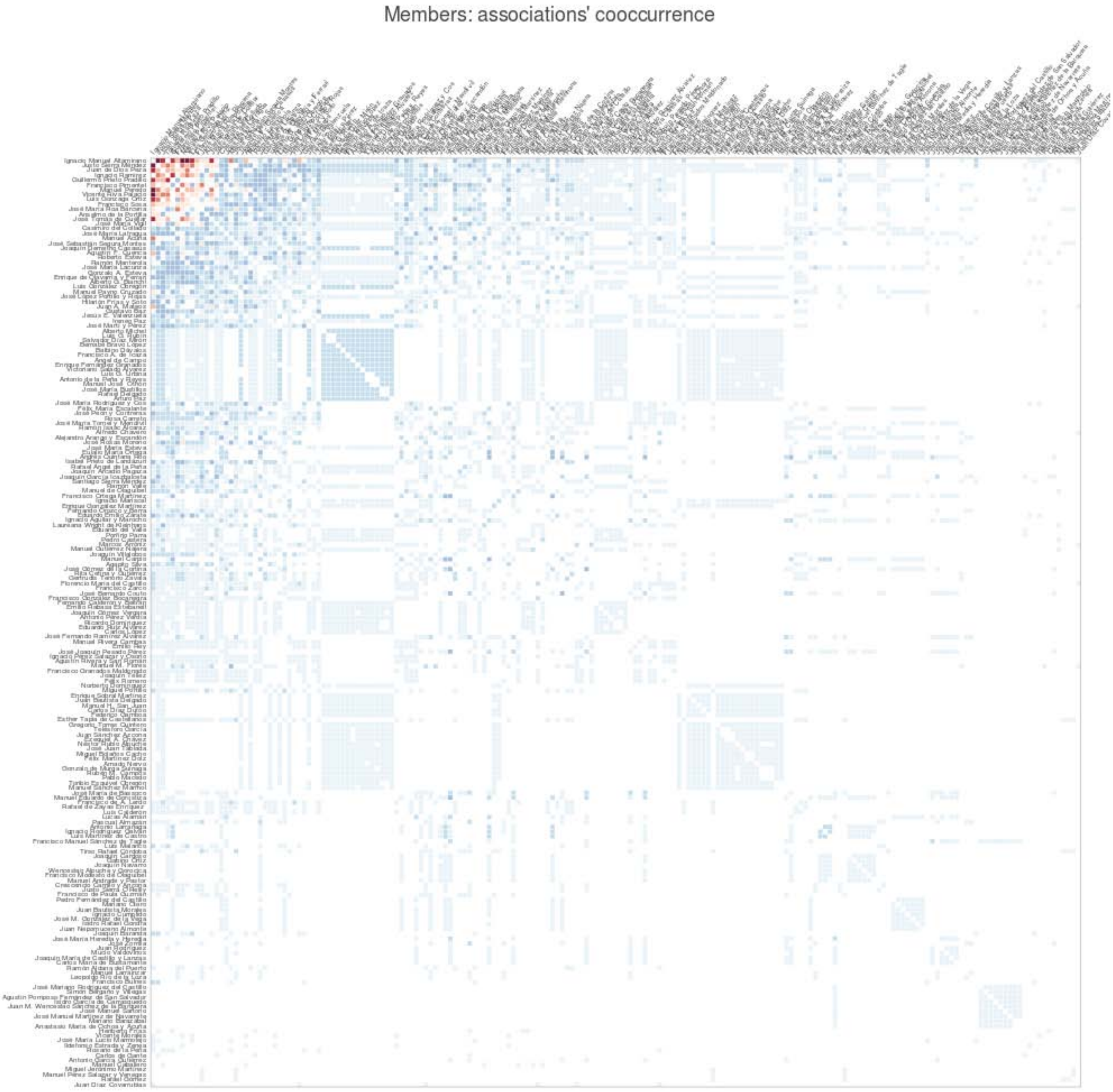


Figure 9. Members' co-occurrence matrix ordered by maximum summed values.

Moreover, when I changed the order of the matrix (by founding date, see Figure 6) I was able to understand these connections in its temporal dimensions. For instance, when zooming on the glyphs for the *Liceo Hidalgo* (see Figure 7) one can easily identify the previous societies with which this association had enough common members to consider them as predecessors, or which other later groups could be considered as successors for the same reasons.

Finally, in another color coding of the glyphs (by the literary movement that was in vogue when these societies were established) I could identify which societies of the same period had more common members (see Figure 8).

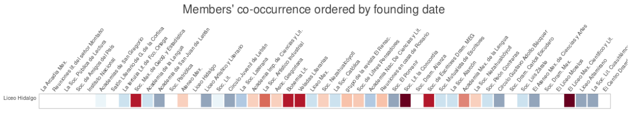


Figure 7. Liceo Hidalgos' co-occurrences, a box-selection of the associations' matrix by founding date

Conversely, I created another matrix –this time comparing members— which was useful to understand which characters co-occurred more often in the same associations and thus address the second question, namely, which were the most recurrent actors in the network (see Figure 9). The result: thirteen characters formed the core of actors who were most involved (see Figure 10). This information, however, could have been obtained with a simple bar-chart. The difference in perspective that this matrix offers is that it allows the user to see that these characters were not only in many but also similar associations (which can be retrieved by hovering the glyphs), and, additionally, it makes evident how proportionally small this core is when compared to the whole matrix.

Finally, to address the second question –namely, the correlation between generations and literary movements–, I created a matrix where associations were ordered by founding date on the y-Axe, and members by birth date on the x-Axe, and where the colors were coded according to their correspondent literary movement (see Figure 11). The dark colors of the glyphs represent the literary movement of a given society (all the blue ones are from the neoclassic movement, for instance), and the light colors in the background represent the members' generations (for example, in light orange -in a vertical division- are all the members of the *Renacimiento* generation). Arranging them like this enabled me to take snapshots of different societies and observe the generations' patterns of membership-adscription. For instance, I could note that although the group formed around the *Renacimiento* magazine was heavily constituted by its homonym generation (see Figure 12), almost half of its members were born in the timeframe of the previous generation (coded with a light yellow background).

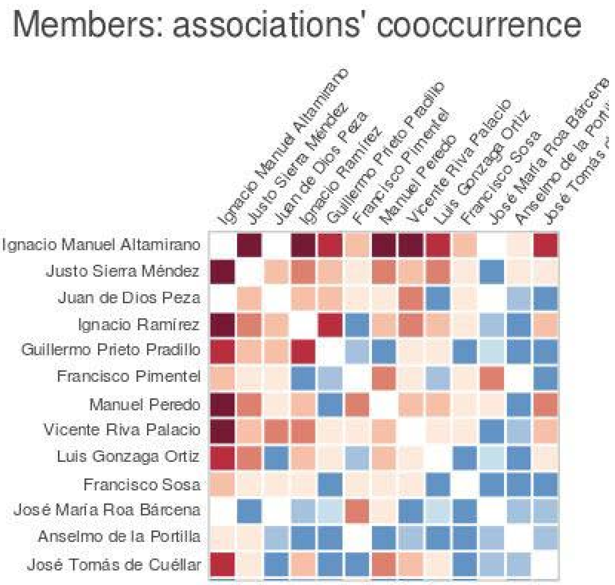


Figure 10. Members' co-occurrence snapshot done with the selection tool of Bokeh's visuals.

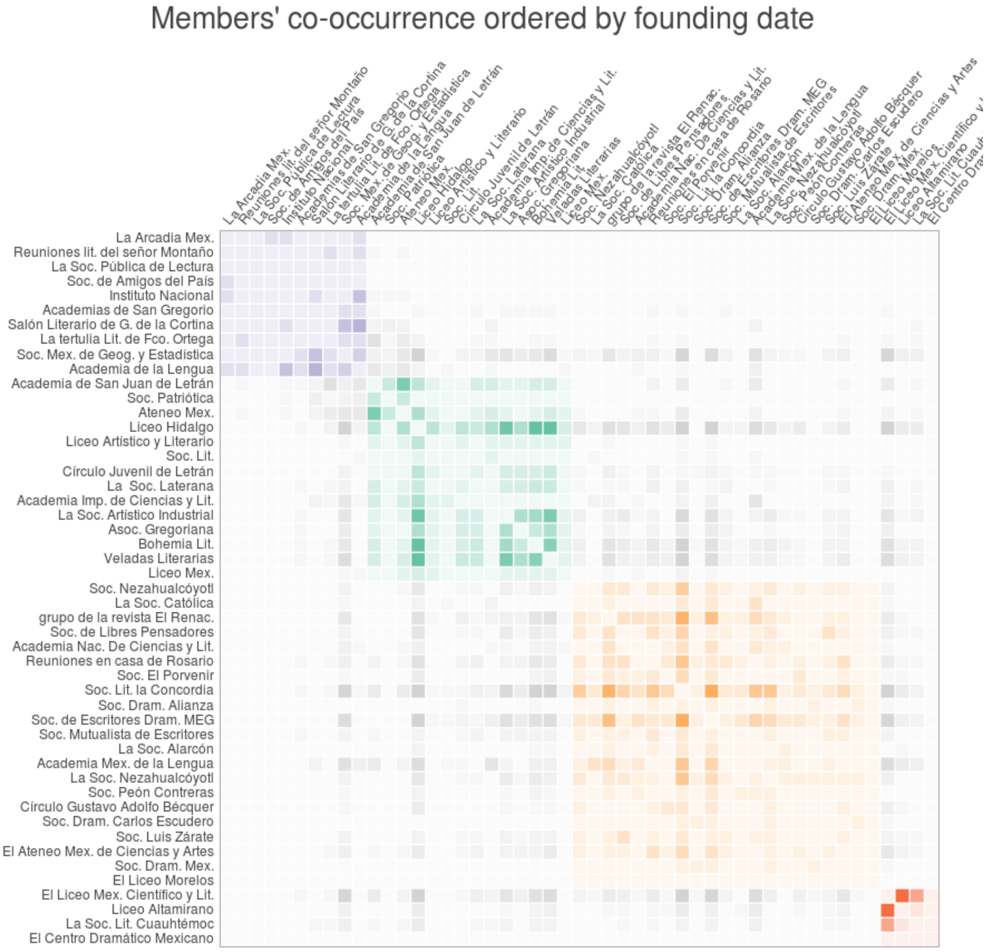


Figure 8. Societies' co-occurrence matrix with literary movements' color-coding.

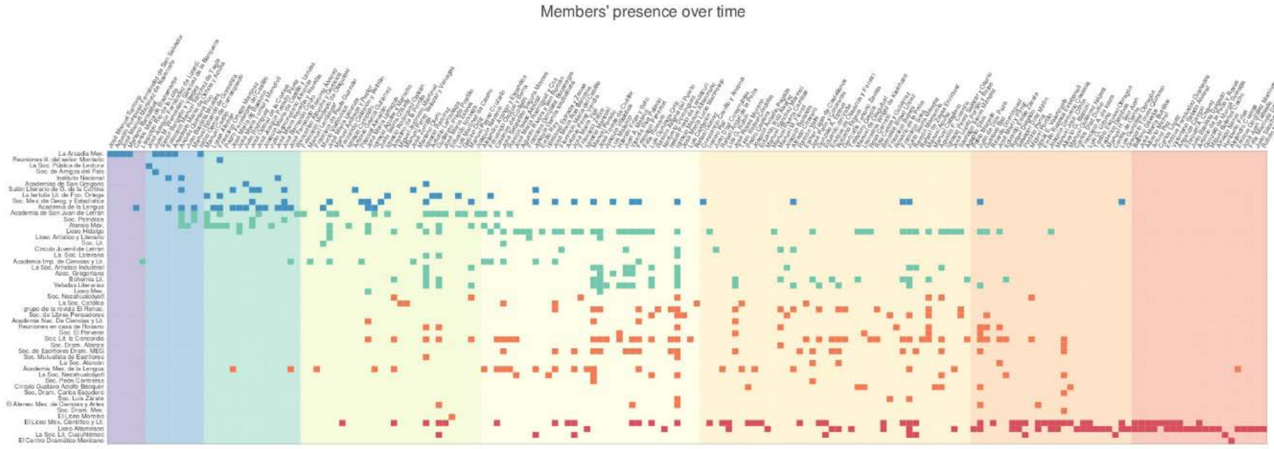


Figure 11. Generations versus literary movements: a co-occurrence matrix

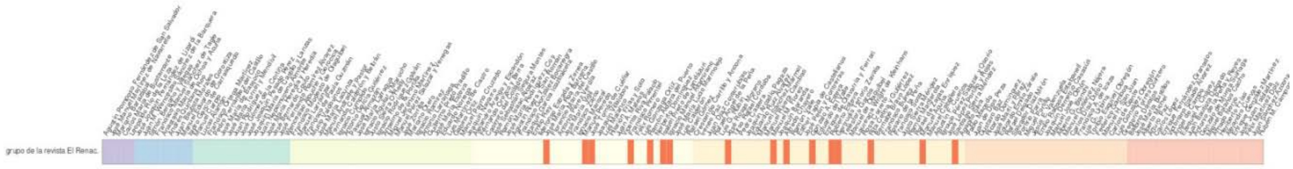


Figure 12. Active members in *Grupo de la Revista el Renacimiento*

Conclusions

In this paper I have shown how customized visualization of modeled data can enable new readings and lead to new understandings of how societies were formed in a key period of national history. Among other things, matrices help us "see" connections between previous categories of literary history (like generations and literary movements), between societies, but also between members, thus supporting new narratives of the lettered city were the alleged homogeneity of this "elite" group can be seen in a nuanced perspective that integrates complexity without sacrificing abstraction.

Bibliography

Chiu, J. (2015). An OpenRefine Reconciliation Service That Queries VIAF. Java https://github.com/codeforkjeff/refine_viaf.

Perales Ojeda, A. (2000). *Asociaciones literarias mexicanas: siglo XIX*. 2nd ed. (Al siglo XIX. Ida y vuelta). México: Universidad Nacional Autónoma de México.

Rama, A. (1984). *La ciudad letrada*. Hanover, N.H., U.S.A.: Ediciones del Norte.

Sánchez, J. (1951). *Academias y sociedades literarias de México*. University of North Carolina.

Spekkink, W. (2014). *Event Graph Layout Wouter Spekkink*. http://www.wouterspekkink.org/?page_id=93 (accessed 20 October 2015).

Verbert, K. V. K. L. (2015). *On the Use of Visualization for the Digital Humanities*. Sydney, Australia http://dh2015.org/abstracts/xml/VERBERT_Katrien_On_the_Use_of_Visualization_for_t/VERBERT_Katrien_On_the_Use_of_Visualization_for_the_Dig.html (accessed 15 December 2015).

Weingart, S. (2015). *Acceptances to Digital Humanities 2015*, (part 2). The Scottbot Irregular http://www.scottbot.net/HIAL/?p=41347 (accessed 23 January 2016).

Wyssen, J. (2014). *How We Created Color Scales*, Website Datavisualization.ch http://datavisualization.ch/inside/how-we-created-color-scales/ (accessed 14 September 2014).

Wyssen, J. (2012). The Salons Project Mapping the Republic of Letters. http://republicofletters.stanford.edu/casestudies/salons.html (accessed 12 November 2014).

Wyssen, J. (2013). Berliner Klassik Geselligkeit-Datenbank Website Berliner Klassik Datenbanken. http://berlinerklassik.bbaw.de/BK/geselligkeit/Suche.html (accessed 25 February 2016).

Project Dialogism: Toward a Computational History of Vocal Diversity in English-Language Literature

Adam Hammond
ahammond@mail.sdsu.edu
San Diego State University, United States of America

Julian Brooke
julian.brooke@gmail.com
University of Melbourne, Australia

Abstract

Over the past several years, we have worked to develop a human-interpretable computational method for quantifying "style" in literary texts. In projects focused on modernist texts, we have demonstrated the usefulness of this approach for studying dialogism (or multi-voicedness) in literature. Now we propose to extend our method to the "big data" scale by using a tool we have created, GutenTag (http://www.projectgutentag.org/). Our research promises insights into the historical evolution of dialogism in English-language fiction.

Aims and Approach

Dialogism — the literary practice of allowing characters to speak in their own distinctive manners, without altering their speech to suit the particular linguistic practices and prejudices of the author — has been recognized as an ethically and politically significant aspect of fiction since the early twentieth century. Russian critic Mikhail Bakhtin, who coined the term "dialogism," has been particularly influential in arguing that the dialogic novel could support pluralistic modes of thought that model democratic social systems. Yet despite the widely recognized importance of dialogism as an analytic category in literary studies, it is one that has proven difficult to study computationally, particularly at the "big data" scale. While style has proven a tractable aspect for computational literary study, and while excellent work has been produced on distinguishing character voices within literary texts using style-based methods (Burrows, 1987; McKenna and Antonia, 1996; Rybicki, 2006), existing approaches present two significant drawbacks. First, their reliance on Burrow's PCA-based methodology means that while this work often produces reliable and insightful results, its computational outputs are generally not human-interpretable; they may be able to show *that* an author distinguishes characters based on linguistic style, but not to tell us *how* they are differentiated. Further, these methods tend not to be suitable to expansion to the "big data" level, since they require significant manual annotation of character speech in the texts under investigation. Our method — a human-interpretable quantitative method for analyzing literary style — and our tool — which performs automatic structural tagging of plain text — make such research possible, and thus open the way for the first large-scale investigation of dialogism in literary fiction.

Background

Since 2011, we have been laying the foundations for a computational history of dialogism in English-language fiction. The first step was developing and refining our six-dimensional approach to quantifying literary style. Our method is based on six discrete aspects of style: objectivity (words that project a sense of disinterested authority); abstractness (words denoting concepts that cannot be described in purely physical terms); literariness; colloquialness; concreteness (words referring to events, objects, or properties in the physical word); and subjectivity (words that are strongly personal or reflect a personal opinion). To build our stylistic lexicons, we produce a relatively small set of words carefully selected for their stylistic diversity, which human annotators evaluate in terms of the six stylistic aspects listed above. Next, we use an automatic procedure to collect information on how these words are employed in all English texts in the 2010 image of *Project Gutenberg* (Brooke et. al., 2016). Using this information, we are able to derive stylistic information for any word in our target text, and to build stylistic profiles for any character or speaking voice within a text.

We have demonstrated the usefulness of our six-style approach to literary problems in two projects. One project focused on free indirect discourse (FID) in Woolf's *To the Lighthouse* and Joyce's "The Dead" (Brooke et. al., 2016). Our intention was to employ our method to test the long-held hypothesis that FID represents a stylistic middle ground between the neutral style of an objective narrator and the more extreme styles of personalized characters as rendered in direct discourse. Our method confirmed this assumption and, because it produces human-interpretable results, shed some new light on Woolf's text in particular, finding that while Woolf's upper-class characters exhibit a conventional power dynamic (they are more authoritative, more literary, more concrete, less objective, and less colloquial than characters of other classes) her female characters reverse these conventional dynamics: they are more objective, more abstract, less colloquial, and less subjective. The other project undertook a quantitative investigation of the problem of voice in T. S. Eliot's *The Waste Land* (Brooke et. al., 2015b). While it is generally agreed that *The Waste Land* is composed of many speaking voices, these voices are not explicitly identified, nor are their points of transition provided. Our work explored methods of automatically segmenting and clustering voices in the text; we used the human-interpretable results of our six-style approach to evaluate the performance of various approaches and arrive at a blended human/machine interpretation.

The other key foundation for our work is GutenTag, an open-source software tool released in October 2015 (Brooke et al., 2015a). Two aspects of GutenTag are particularly relevant to the present project. First, it allows users to quickly create large, customized literary corpora. Working from the 2010 image of Project Gutenberg (PG), metadata provided by PG and derived automatically from other sources, and our automatic decision-tree genre classifier, one can, for example, easily assemble a large corpus of nineteenth-century novels in English. Second, it uses our sophisticated rule-based method to automatically generate reliable, genre-specific structural tags in standard TEI XML. Crucially, our tagging system is able to distinguish character speech from narration and identify individual characters in novels; and to separate character speech from stage directions and setting descriptions in plays, associating each speech with a character in the dramatis personae.

Developing a Metric for Dialogism

With the major pieces in place to conduct our research into the history of dialogism, our main task is to develop a reliable quantitative metric for the dialogism of a given literary text. We will proceed by calculating stylistic profiles for each character using methods already established in previous work. We will include a minimum word cutoff to exclude characters for which the stylistic profile is likely to be too noisy due to lack of data. Next, assuming multiple characters, for each style we will treat each character as a datapoint and calculate a weighted variance, where weights are applied based on the relative proportion of speech by each character, to produce a number that indicates overall stylistic variation across characters for that stylistic dimension in the given text. We can average across styles to produce a single metric.

We will experiment with calculating dialogism based on stylistic variance between (a) individual characters, (b) groupings of characters (based on gender, social background, age, etc.), and (c) social networks. This will allow us to track (a) texts in which the speech of individual characters is highly varied, (b) texts in which, for example, male characters speak in a highly distinct manner from female characters, and (c) texts in which members of distant branches of a derived social network speak in their own differentiated styles. To carry out this analysis, certain modifications will be necessary to the GutenTag system; namely, methods to detect social networks in texts and to automatically identify character groupings (a method for identifying character gender is already in place; methods for identifying other groups will be more difficult, but we will explore them).

Research Questions

With these technical foundations in place, we will be ready to present our answers to some of the following questions. Our aim is not to cover all of them, but to offer detailed investigations of those that yield the most interesting results.

- Which texts in PG are the most stylistically diverse, according to our quantitative definition? Are they texts by authors traditionally celebrated as dialogic (Austen, Woolf, Joyce, translations of Dostoevsky) or are they non-canonical texts? If the latter, do these texts register qualitatively as "dialogic" to a human reader? What does a close reading of these texts reveal about the viability of our automatic method?
- How does stylistic diversity map onto historical time? Do periods of political turmoil (wars, revolutions, natural disasters, strikes, etc.) correspond to changes in the average stylistic diversity of fiction? What can we learn about the social role of fiction by studying this relationship?
- How is stylistic diversity distributed geographically? Which regions produce the most stylistically varied writing?
- How is stylistic diversity distributed among genres? Does our method support or refute Bakhtin's claim that the novel is the most dialogic of the genres, and poetry the least dialogic? Can dialogism be meaningfully compared across genres?
- Are authors of different ages, classes, and genders more or less likely to produce dialogic fiction?

How does the stylistic diversity of fiction track against that of non-fiction? Does non-fiction become more or less dialogic over time, and does it follow a similar curve to that of fiction? Do changes in the dialogism of fiction anticipate changes in non-fiction, or are the two unrelated?

Bibliography

Brooke, J., Hammond, A., and Hirst, G. (2016). Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction. *Digital Scholarship in the Humanities*, **2**(2): 1–17.

Brooke, J., Hammond, A., and Hirst, G. (2015a). GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. *Workshop on Computational Linguistics for Literature.* Denver: NAACL, pp. 1–6.

Brooke, J., Hammond, A., and Hirst, G. (2015b). Distinguishing Voices in *The Waste Land* Using Computational Stylistics. *Linguistic Issues in Language Technology*, **12**(2): 1–43.

Burrows, J .F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

McKenna, C. W. F., and Antonia, A. (1996). 'A Few Simple Words' of Interior Monologue in *Ulysses*: Reconfiguring the Evidence. *Literary and Linguistic Computing*, **11**(2): 55–66.

Rybicki, J. (2006). Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and Its Two English Translations. *Literary and Linguistic Computing*, **21**(1): 91–103.

Measuring the Dynamics of Lexico-Semantic Change Since the German Romantic Period

Johannes Hellrich
johannes.hellrich@uni-jena.de
Research Training Group "The Romantic Model. Variation - Scope - Relevance", Friedrich-Schiller-Universität Jena, Jena, Germany

Udo Hahn
udo.hahn@uni-jena.de
Jena University Language and Information Engineering (JULIE Lab), Friedrich-Schiller-Universität Jena, Jena, Germany

Introduction

The dynamics of language change over time are most evident in the lexicon component of natural languages. In particular, the gradual semantic changes words may undergo have a strong effect on the comprehension of historical texts by modern readers. Yet, efforts to automatically detect and trace this lexical evolution are scarce. Our study follows the work of Kim et al. (2014) who detected lexico-semantic changes in English texts over the 20th century via a series of neural network language models. Our models were trained on the German part of the *Google Books Ngram*[1] corpus (Michel, et al., 2011; Lin et al., 2012), which covers over 657k German books. Such models have the particular advantage that they can be queried for the semantic similarity of arbitrary words. We tested this query option by sampling nouns from *Des Knaben Wunderhorn* (Arnim and Brentano, 1806-1808), a collection of German folk poems and songs from the German Romantic period. The choice of this volume is merely motivated by our interest in the literary period it belongs to. We detected interesting semantic changes between 1798 (often taken as the starting point for the German Romantic period) and 2009 (last year in the Google corpus).

Methods

Using the specific contexts in which words appear in order to determine the (distributional) meaning of words is an old idea from linguistic structuralism (Firth, 1957). For a long time, this appealing approach could not have been seriously investigated due to the lack of suitably large corpora and adequate computational power to deal with distributional patterns of words on a larger scale. Thus, only few studies on automatically detecting semantic change have been conducted up until now, with a clear focus on the high-volume data provided by Google Books. This collection is widely popular due to its immediate availability and enormous coverage despite well-known problems stemming from both the quality of optical character recognition (OCR) and the sampling strategies used to compile it (Pechenick et al., 2015).[2]

Early approaches towards modeling lexico-semantic change patterns used frequency and bi-gram co-occurrence data (Gulordava and Baroni, 2011), as well as (context-based) classifiers (Mihalcea and Nastase, 2012). Riedl et al. (2014) built distributional thesauri to cluster similar word senses. All of these approaches detected lexico-semantic changes between multiple pre-determined periods. In contrast, neural network language models can be used to detect changes between arbitrary points in time, thus offering a longitudinal perspective (Kim et al., 2014; Kulkarni et al., 2015). In our experiments, we use a skip-gram model, a simplified neural network that is trained to predict plausible contexts for a given word, thereby generating (computationally less expensive) low-dimensional vector space representations of a lexicon (Mikolov et al., 2013). Despite their simplicity, neural network language models are a state-of-the-art approach, with details concerning ideal implementation solutions and training scenarios still being under dispute (Baroni et al., 2014; Schnabel et al., 2015).

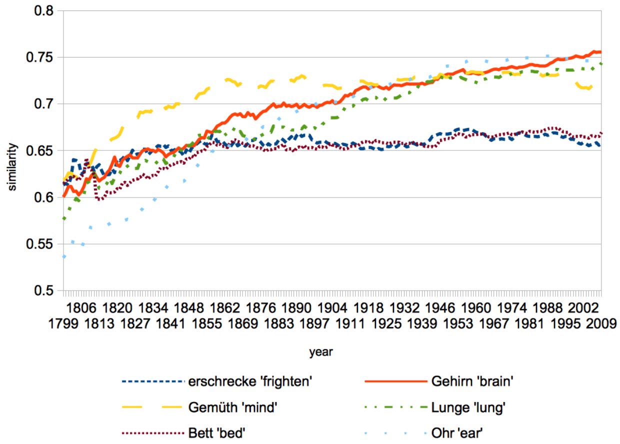
Experiment

We trained our models on 5-grams spanning the years 1748 to 2009, using a uniform sampling size of 1M 5-grams per year; the first 50 years were used for initialization only. Test words for high-lighting semantic change patterns were selected from *Des Knaben Wunderhorn* by identifying the ten most frequent nouns, i.e. *Gott* [‘god’], *Herr* [‘lord, mister’], *Liebe* [‘love’], *Tag* [‘day’], *Frau* [‘woman, miss’], *Mutter* [‘mother’], *Herz* [‘heart’], *Wein* [‘wine’], *Nacht* [‘night’] and *Mann* [‘man’]. For each of these ten nouns we selected the three words most similar to them (according to the cosine of their respective vector representations) during 1799 and 1808 and between 2000 and 2009, tracking how the similarity of these words developed between 1798 and 2009. The programs used for our experiments and resulting data are publically available via GitHub.[3]

Results

The cosine similarity between the 1798 and the 2009 vector representation of the ten test words is rather high, ranging from 0.72 for *Mann* to 0.84 for *Wein*, thus showing only minor semantic changes. Manual interpretation of their most similar words revealed an interesting change for *Herz* (see Fig. 1) that is nowadays more similar to other anatomical terms (such as *Gehirn* [‘brain’], *Lunge* [‘lung’], or *Ohr* [‘ear’]) and less likely to be used metaphorically (such as indicated by *erschrecke* [‘frighten’], or *Gemüth* [archaic for ‘mind’]). As this change predates Google Books’ tendency to overrepresent scientific texts (at least for English, cf. Pechenick et al., 2015) this finding can be assumed to be an example of true lexico-semantic change. The example also demonstrates a need for a metric incorporating frequency information and normalization of input, since *Gemüth* is an archaic form for *Gemüt* nonconformant with modern German spelling conventions, although it is rated as currently similar to *Herz*.

Fig. 1 Lexical semantics of *Herz* [‘heart’] as expressed by its similarity with six other words; similarity-axis not depicting whole range of possible values (0–1)



Conclusion

This research note has gathered preliminary evidence for the feasibility of corpus-driven studies into German diachronic semantics. We advocate a computational, neural network-based approach where the evolution of lexico-semantic changes is traced by similarities of distributional patterns in the context of words over time.

Looking backwards for semantic changes is, however, constrained by the quality and quantity of linguistic data available. While the primary corpus we use for determining semantic evolution patterns, the Google Books Ngram corpus, is remarkably large, it suffers from a idiosyncratic sampling policy, as well as OCR shortcomings and even more advanced issues, such as the absent normalization of historic orthographic variants. Other historic corpora dealing with the latter quality issues (such as the *Deutsches Textarchiv*) are plagued by their comparatively minuscule size.

Future research in Digital Humanities, besides dealing with these issues, will exploit the similarity data in order to make proper use of them under a humanities' perspective and, thus, hopefully determine the added value of such computational results. This can be achieved by incorporating complementary types of data (e.g. historical, economic ones) to render additional evidences to change patterns. Since this is a huge and complex task, we plan to make our similarity data publically available on a website, together with an easy-to-use interface, as a humanities tool for comparative, diachronic lexico-semantic studies, with several user-adjustable parameters (e.g. different grain sizes of time intervals, alternative ranking metrics, *etc.*). From a methodological perspective, we plan to focus our research on protocols for training models covering long timespans, a metric to measure the quality of historic language models (probably including the need for a manual evaluation) and a way to include frequency information–a word which is no longer used cannot be said to be unchanged in its semantics. Such a system would ideally be tested by an in-depth study of the semantics of carefully selected words, including a comparison with prior, hermeneutically guided work in the humanities as a rich, yet completely informal background theory.

Funding

This work was supported by the DFG-founded Research Training Group "The Romantic Model. Variation - Scope - Relevance" [grant GRK 2041/1].

Bibliography

Arnim, A. von and Brentano, C. (1806-1808). *Des Knaben Wunderhorn* **1**(3), (Annotated TCF version provided by the Deutsches Textarchiv).

Baroni, M., Dinu, G. and Kruszewski, G. (2015). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*,**1**: 238–47.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, pp. 1–32.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus .*Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics @EMNLP 2011*, pp. 67–71.

Kim, Y., et al. (2014). Temporal analysis of language through neural language models . *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 61–65.

Kulkarni, V., et al. (2015). Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*, pp. 625–35.

Lin, Y., et al. (2012). Syntactic annotations for the Google Books Ngram Corpus . *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 169–74.

Michel, J.B., et al. (2011). Quantitative analysis of culture using millions of digitized books . *Science*, **331**(6014): 176–82.

Mihalcea, R. and Nastase, V. (2012). Word epoch disambiguation: Finding how words change over time . *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, **2**: 259–63.

Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26 (NIPS2013),* pp. 3111–119.

Pechenick, E. A., Danforth, C. M. and Dodds, P. S. (2015).

Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* **10**(10): e0137041.

Riedl, M., Steuer, R. and Biemann, C. (2014). Distributed distributional similarities of Google Books over the centuries . *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1401–405.

Schnabel, T., et al. (2015). Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP '15)*, pp. 298–307.

Notes

[1] An *n*-gram is a sequence of *n* words plus information on their frequency/probability of occurrence for a given corpus. The available version of the corpus does not consist of running text, but of *n*-grams instead.

[2] The *Deutsches Textarchiv* (DTA) can be considered as a counter example, at least, as far as the quality of OCR is concerned. Yet, DTA suffers from tremendous size limitations in comparison with the (German portion of the) Google corpus, since this corpus for historic German texts contains only about 2.4k texts (http://www.deutschestextarchiv.de/list).

[3] https://github.com/hellrich/dh2016

Academic Pillow-Talk and Two Immersive Explorations of Linguistic Space

Rachel Marion Hendery
r.hendery@uws.edu.au
Western Sydney University, Australia

As Virtual Reality emerges as an accessible technology, researchers have begun to experiment with its potential for data visualisation. In this paper I will use two particular VR visualisations of linguistic data as a springboard for discussing the affordances of different immersive technologies for research.[1]

The first visualisation is a geographically-anchored walkthrough of data held in the PARADISEC linguistic archive (Thieberger and Barwick, 2012). A user moves across a representation of a geographic region and sees markers on the landscape representing the metadata of the relevant PARADISEC materials for that location: number of speakers, amount and diversity of material held. Audio and text appear when the user gazes at a marker. Looking up, the user can also see the historical relationships between the languages as a network. Such a visualisation could be adapted to represent the holdings of other kinds of archives, so the discussion in this paper has implications for digital humanities more generally.

The second visualisation is a more abstract three-dimensional cloud of coloured points laid out on three axes. The user is located initially in the midst of this cloud, and can move through it in any direction. The colours, sizes and shapes of the points, as well as their location on the x, y, and z axes, can be mapped to variables of any kind of data, and therefore this too has implications beyond linguistic research. For my purposes, they represent linguistic features derived from the *World Atlas of Language Structures* (WALS) (Dryer and Haspelmath, 2013). To demonstrate one use case, different measures of linguistic complexity can be graphed on the x, y, and z axes (e.g. WALS features 1A+2A, 22A, 49A, etc). Colouring the points according to the geographical location of the languages, with shape determined by genetic affiliation makes it easy to explore clusters of types of complexity.

Both of these visualisations can be viewed in a pseudo-3D environment using WebGL in a desktop browser. For a more immersive experience, however, they can also be viewed inside a virtual reality headset such as the Google Cardboard or Oculus Rift. Alternatively, the visualisations can be displayed on the inside of a fulldome such as those used in a planetarium. Through a recent successful LIEF grant, Western Sydney University and a number of partners have been able to construct an ultra-high resolution experimental fulldome ('DomeLab' http://www.niea.unsw.edu.au/research/projects/domelab).

Display via virtual reality headset is the most straightforward sort of immersion. The viewer is embodied amidst the data, and simple movements from everyday life (e.g. looking around) translate directly into the virtual space. Display of the visualisation in a browser is at the other extreme of (non)immersion. However, many users are familiar with the translation of three-dimensional space onto a flat screen from modern computer games, so it can still give a sense of movement, rotation, and interactivity.

A hybrid of these two experiences, perhaps less familiar, is the dome. The user lies on the floor inside the dome, and the visualisation is displayed on the concave walls of the dome above. Because the user is surrounded by the screen, the experience is far more immersive than the display of a visualisation in a computer browser. Metaphors of movement and space have to be adapted for such a display, however, as most of the display surface is above the user.

This paper will discuss the reasons why we might want to explore linguistic data visually, and what we gain or lose by doing so in the various environments described above. Lev Manovich notes that information visualisation is the representation of datasets in such a way as to reveal structure (Manovich, 2010). Linguists are obsessed with structure. Since at least Pāṇini (4th century BC, see Vasu, 1891), linguists have conceived individual languages as highly structured; since the Neogrammarians if not before, 'language' as an abstraction that changes and exists outside of its speakers has also been considered a highly structured object of study; sociolinguists are interested in the structure of speech communities and speaker networks; psycholinguists and neurolinguists in the linguistic structures of the mind and the brain. Linguistics is therefore a field that is primed to search for and find structure in its data, and is constantly searching for new ways to do so.

Manovich also describes data visualisation as a matter of reduction and spatialisation, often involving a remapping of the non-visual to the visual. For linguists, reducing linguistic data to features and remapping it to the visual (audio to text, for example) are business as usual. What I would like to discuss in this paper is the spatialisation of the data, and the embodiment of the researcher(s) within that data space.

A very active branch of linguistics involved in data visualisation is linguistic typology. Typology concerns itself with how languages are distributed within the possibility space of all conceivable languages: i.e. "what's where and why" (Bickel, 2007). Because linguistic data has an inherent geographical dimension, the predominant linguistic visualisation typologists use is a map. I have followed this lead for my PARADISEC visualisation. However, Manovich also points out that both designers and their audience tend to treat spatial dimensions as primary (Manovich, 2010:8). Therefore it makes sense also to experiment with mapping these to other, perhaps more significant features of the linguistic data. This is the impetus behind my data graph visualisation. By translating the WALS data to a more abstract 3D graph, different relationships, clusters and structures become apparent. In the full paper I demonstrate examples of this.

Data visualisations can have a variety of purposes (see e.g. Dransch, 2000; Purchase et al., 2008). They can be a way to explore data to generate ideas or observations that then inspire future research. They can be an attempt at scientific modelling, instantiating more fully formed pre-existing ideas. Or they can be a way of communicating information to others. An early study of user experience in 2D, 3D and immersive data visualisation (Modjeska, 2000) showed that users generally find 2D representations of the data more efficient, but enjoyment and motivation increases with the degree of immersiveness. This suggests that VR visualisations might be best suited for the playful exploration of data described above than for serious scientific modelling. They may also be well suited for communication of ideas when it is not essential that the audience grasp complex details.

These are therefore the motivations behind the two linguistic data visualisations I describe. With the PARADISEC map visualisation it is not essential that a user come away with perfect recall of what exactly the archive holds, but rather with a sense of its richness, and increased motivation for using the archive in the future. The data graph visualisation is aimed at researchers who want to explore language data creatively in order to generate new ideas.

Finally, I discuss how the affordances of the VR headset versus the Dome depend on one's beliefs about the locus of knowledge production. In the 'lone wolf' researcher model, an individual generates ideas in a relative vacuum, or at least, the important connections for the researcher are among and with the data itself, not with other people. For this model, a VR headset is ideal. In the world of the headset, the researcher is alone with the data: the mundane is quite literally blacked out. The dome, on the other hand, is designed to be experienced communally. The floor is covered with large pillows, and people lie head-to-head or side-by-side. Curtains delimit the boundaries, but this threhold is permeable: people come and go. It is natural to discuss the visualisations above with those who lie beside you in the half-dark. In this way, research conversations become a kind of academic pillow-talk, naturally imbued with the playfulness, intimacy, and gentleness of such. In the VR headset, you become part of the data; in the dome you become part of a community.

Bibliography

Bickel, B. (2007). Typology in the 21st century: major current developments. *Linguistic Typology*, **11**(1): 239-51.

Dransch, D. (2000). The use of different media in visualizing spatial data. *Computers & Geosciences*, **26**(1): 5-9.

Dryer, M. and Haspelmath, M. (Eds.) (2013). *The World Atlas of Language Structures Online*. MPI for Evolutionary Anthropology. http://wals.info

Manovich, L. (2010). What is visualization? *PAJ: The Journal of the Initiative for Digital Humanities, Media, and Culture*, **2**(1).

Modjeska, D. (2000). *Hierarchical data visualization in desktop virtual reality.* Doctoral dissertation, University of Toronto.

Purchase, H., et al. (2008). Theoretical foundations of information visualization. In Kerren, A., Stasko, John T., Fekete, J-D. and North, C. (Eds), *Information Visualization.* Springer, pp. 46-64.

Thieberger, N., and Barwick, L. (2012). Keeping records of language diversity in Melanesia, the Pacific and regional archive for digital sources in endangered cultures (PARADISEC). *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*, pp. 239-53.

Vasu, S. C. (1891).*The Ashtadhyayi of Panini* vol. 1, Motilal Banarsidass, Delhi, India.

Notes

[1] My development of the linguistic visualisations discussed here was funded by a transdisciplinary grant from the Centre of Excellence for the Dynamics of Language (COEDL).

Tracing the Genesis of Pessoa's Envisioned Work: a Digital Edition of his Editorial Projects and Publications

Ulrike Henny
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Germany

Pedro Sepúlveda
psepulveda@fcsh.unl.pt
Universidade Nova de Lisboa, Portugal

Introduction

A digital edition of Fernando Pessoa, considered one of the most significant Modernist poets and writers, is being established through a collaboration between scholars from the Institute of Literature and Tradition (IELT) of the New University of Lisbon and the Cologne Center for eHumanities (CCeH) of the University of Cologne. The edition focuses on the contrast between the potential character of Pessoa's numerous lists of editorial projects and his few actual publications in lifetime.

Though the digital platform is primarily designed to support the research aims of the project, the procedures can be reused by others (encoding, scripts). From a theoretical standpoint, the question of „work genesis" (on a macro and micro level) is relevant to a wider audience. In the case of Pessoa, the publication of the editorial lists and plans together with the published works in lifetime contributes to the discussion on the genesis and status of each stage of a work.

State of Research and Aims

It has been much debated by literary scholars whether Fernando Pessoa's work should be characterized as fragmentary or unitary (see Coelho, 1949; Gusmão, 2003; Martins, 2003; Sepúlveda, 2013). Due to the low number of publications in lifetime, the existence of a vast literary archive and the use of many fictional authors, named "heteronyms" by the author, the coherence of his work has been questioned. The present edition, in contrast, is based on the assumption that Fernando Pessoa constantly worked on the organization of his work, as witnessed by the many editorial lists, notes and plans which are part of his estate (Cunha, 1987; Sepúlveda, 2013).

These papers are the subject of the digital edition at hand, together with the 60 publications of poems that Pessoa realized in journals and literary magazines. That way, the genesis and evolution of his work as planned by himself can be studied and the published texts be examined in the light of the editorial projects preceding, accompanying and following them during the author's life. A significant part of the documents witnessing the editorial activity of the author have not been published before. For those which have been published, past editions in bookform define themselves by following either the first or the last version of a text, or by choosing from the existent variants from a hermeneutical standpoint (Duarte, 1988; Galhoz, 1993; Martins, 2011; Castro, 2013).

As the textual variants and hesitations are critical to understand shifts in the conception of the work, e. g. when work titles are assigned to heteronyms or the composition of a planned publication is changed, the digital edition aims at offering several coexistent forms of transcription. To facilitate the understanding of the texts and the work they trace, editorial comments are to be added to the documents.

In terms of editorial methodology, the digital edition targets at exploring the possibilities to combine different editorial procedures from a documentary, a diplomatic, a genetic and an enriched edition, overcoming previous oppositions between different editorial approaches to the poet's work. Relations between the editorial projects and the published texts are formalized via the encoding of "work references". By extending the concept of work to include projections which the author makes on a macro level, i.e. outlining his works in terms of titles, authorship, structure and publication organs, the present edition contributes to the debate about the status of the work and its place in the digital edition (Robinson, 2013; Sahle, 2013).

Methods and Results

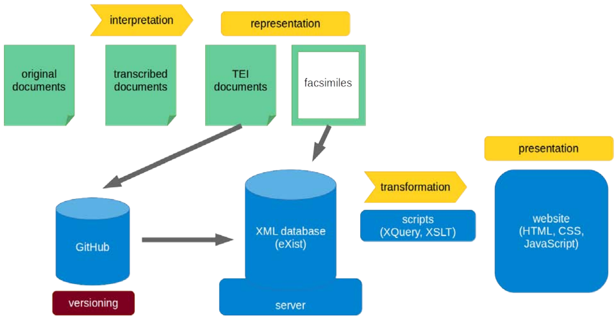
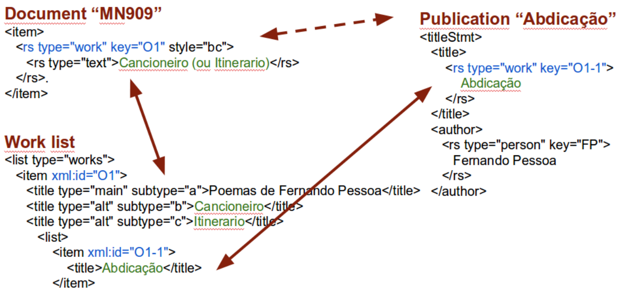


Figure 1: Technical setup and workflow

The realization of the digital edition relies primarily on procedures and technologies coming from the X universe. Transcriptions of the documents and metadata are encoded according to the standard of the Text Encoding Initiative (TEI-P5). The TEI documents are stored in an XML database (eXist) hosted on a web server together with the digital facsimiles. While the work on the documents is going on, a GitHub repository reachable at < https://github.com/cceh/pessoa> is used to hold successive versions of both the TEI documents and the application which is developed. Scripts written in XQuery and XSLT control the transformation of the documents' underlying representation into the presentation layer. Some of the components shipped with eXist are used: Bootstrap, JQuery and eXist's Templating Framework plus a SIMILE timeline and the OpenSeadragon image viewer.

The TEI encoding is controlled with a RelaxNG schema based on an ODD file. The recorded metadata include bibliographic details, the date or period of the document's genesis or publication and a genre classification (editorial notes, editorial lists, editorial plans, poetry).

Among the encoded textual phenomena are additions, deletions, substitutions, omissions, variants and notes. First and last variants are especially marked to allow for the establishment of a first and last version of the text. Moreover, the following types of entities are labeled: persons, journals, texts and works. Special attention is paid to the four principal authorial figures of Pessoa: Alberto Caeiro, Álvaro de Campos, Fernando Pessoa and Ricardo Reis. Besides registering their mere occurrence, the role they have is determined (they can be mentioned as author, editor, translator or topic of a work). Occurring persons, journals and works are managed in a central knowledge base which serves as the basis for their identification in the documents and the creation of comprehensive indexes. The following figure illustrates how the relationship between an editorial list and a publication is established via the reference to works:



Document "MN909"

```
<item>
  <rs type="work" key="O1" style="bc">
    <rs type="text">Cancioneiro (ou Itinerario)</rs>
  </rs>.
</item>
```

Work list

```
<list type="works">
  <item xml:id="O1">
    <title type="main" subtype="a">Poemas de Fernando Pessoa</title>
    <title type="alt" subtype="b">Cancioneiro</title>
    <title type="alt" subtype="c">Itinerario</title>
      <list>
        <item xml:id="O1-1">
          <title>Abdicação</title>
        </item>
```

Publication "Abdicação"

```
<titleStmt>
  <title>
    <rs type="work" key="O1-1">
      Abdicação
    </rs>
  </title>
  <author>
    <rs type="person" key="FP">
      Fernando Pessoa
    </rs>
  </author>
```

Figure 2: Example of the encoding of work references in TEI

The document MN909 is an editorial list. One item of the list is a mention of the work "Cancioneiro" or "Itinerario". In the central work list, those titles are assigned to the poems of Fernando Pessoa. One of the poems is "Abdicação" which in turn is represented by a publication carrying that title. Thus, the relationship between the editorial list and the publication is made explicit through the encoding and can be exploited e.g. in the edition's presentation.

Work on the digital edition has been going on since October 2014. It is planned to launch the edition within 2016. Until then, more documents will be uploaded and work on the presentation will be finalized. Results so far can be viewed at < http://projects.cceh.uni-koeln.de/pessoa>.

The presentation is multilingual (Portuguese and English). Access to the material is facilitated by different browsing options: by author, documents, publications, works, genre, and chronology. A simple and an advanced search enable the user to make specific and custom requests. As a first analytic approach to the question of the evolution of Pessoa's work during lifetime, an interactive timeline visualising the chronology of editorial projects and publications has been created.

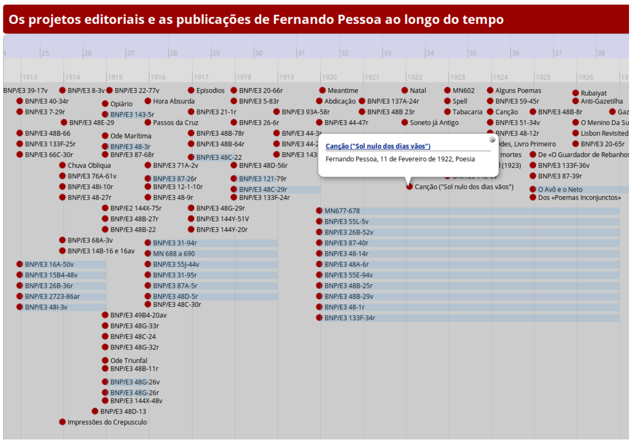
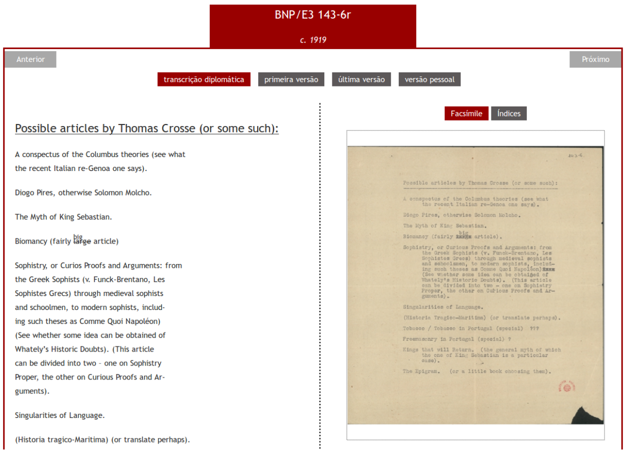


Figure 3: Timeline of Pessoa's editorial projects and publications <http://projects.cceh.uni-koeln.de/pessoa/pt/timeline.html>

The publications and documents are presented in a synoptic view, juxtaposing the transcription on the left and the facsimile on the right side. The presentation of the documents containing the editorial lists, notes and plans is enhanced by offering the possibility to switch between the diplomatic transcription, the critically established text in the first and last version and a so-called "customized version" where the user can combine transcription features (e. g. if original line breaks should be maintained or omitted). Additionally, the synoptic view can be changed to transcription-index mode, where occurring persons, texts and journals are shown alongside the transcription of the document and links to the central indexes are established. The following figures illustrate the above mentioned features:



Possible articles by Thomas Crosse (or some such):

A conspectus of the Columbus theories (see what the recent Italian re-Genoa one says).

Diogo Pires, otherwise Solomon Molcho.

The Myth of King Sebastian.

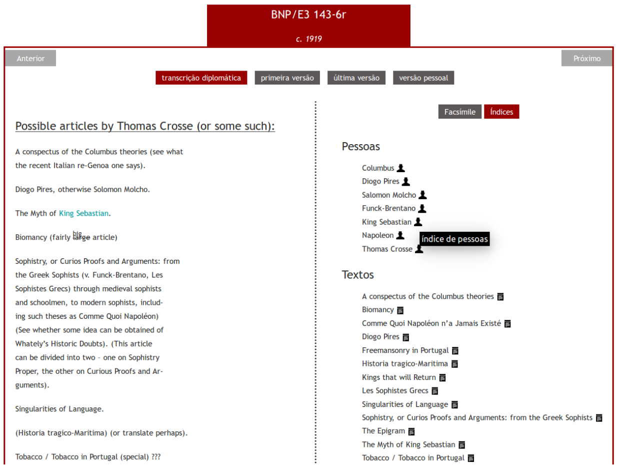
Biomancy (fairly ~~large~~ big article)

Sophistry, or Curios Proofs and Arguments: from the Greek Sophists (v. Funck-Brentano, Les Sophistes Grecs) through medieval sophists and schoolmen, to modern sophists, including such theses as Comme Quoi Napoléon) (See whether some idea can be obtained of Whately's Historic Doubts). (This article can be divided into two - one on Sophistry Proper, the other on Curious Proofs and Arguments).

Singularities of Language.

(Historia tragico-Maritima) (or translate perhaps).

Figure 4: Synopsis of editorial list and facsimile <http://projects.cceh.uni-koeln.de/pessoa/BNP_E3_143-6r#facsimile>



Possible articles by Thomas Crosse (or some such):

A conspectus of the Columbus theories (see what the recent Italian re-Genoa one says).

Diogo Pires, otherwise Solomon Molcho.

The Myth of King Sebastian.

Biomancy (fairly big ~~large~~ article)

Sophistry, or Curios Proofs and Arguments: from the Greek Sophists (v. Funck-Brentano, Les Sophistes Grecs) through medieval sophists and schoolmen, to modern sophists, including such theses as Comme Quoi Napoléon) (See whether some idea can be obtained of Whately's Historic Doubts). (This article can be divided into two - one on Sophistry Proper, the other on Curious Proofs and Arguments).

Singularities of Language.

(Historia tragico-Maritima) (or translate perhaps).

Tobacco / Tobacco in Portugal (special) ???

Pessoas

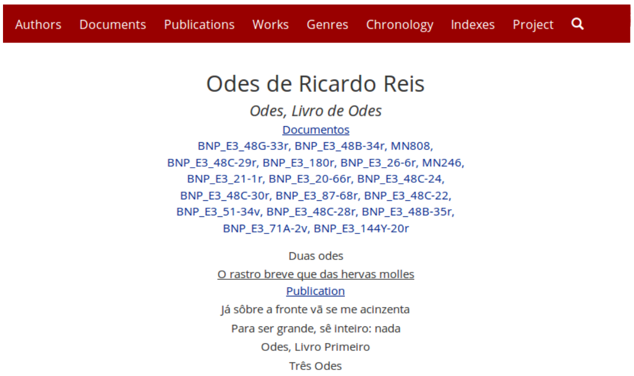
Columbus
Diogo Pires
Salomon Molcho
Funck-Brentano
King Sebastian
Napoleon
Thomas Crosse

Textos

A conspectus of the Columbus theories
Biomancy
Comme Quoi Napoléon n'a Jamais Existé
Diogo Pires
Freemansonry in Portugal
Historia tragico-Maritima
Kings that will Return
Les Sophistes Grecs
Singularities of Language
Sophistry, or Curios Proofs and Arguments: from the Greek Sophists
The Epigram
The Myth of King Sebastian
Tobacco / Tobacco in Portugal

Figure 5: Synopsis of editorial list and index references <http://projects.cceh.uni-koeln.de/pessoa/BNP_E3_143-6r#index>

For each of the four principal authorial figures, the relationships between "works", "publications" and "documents" are presented on an individual page (see figure 6). In the example, the works of Ricardo Reis are listed starting with attested alternative titles of the whole work ("Odes de Ricardo Reis", "Odes", "Livro de Odes"), followed by the titles of individual works. Links to the documents show where the works are mentioned in the editorial projects and links to publications indicate where a work has been published by Pessoa.



Odes de Ricardo Reis

Odes, Livro de Odes

Documentos

BNP_E3_48G-33r, BNP_E3_48B-34r, MN808, BNP_E3_48C-29r, BNP_E3_180r, BNP_E3_26-6r, MN246, BNP_E3_21-1r, BNP_E3_20-66r, BNP_E3_48C-24, BNP_E3_48C-30r, BNP_E3_87-68r, BNP_E3_48C-22, BNP_E3_51-34v, BNP_E3_48C-28r, BNP_E3_48B-35r, BNP_E3_71A-2v, BNP_E3_144Y-20r

Duas odes
O rastro breve que das hervas molles
Publication
Já sôbre a fronte vã se me acinzenta
Para ser grande, sê inteiro: nada
Odes, Livro Primeiro
Três Odes

Figure 6: Work list for Ricardo Reis <http://projects.cceh.uni-koeln.de/pessoa/obras/O3>

Conclusion

The digital edition of Fernando Pessoa shows that the author was continuously concerned with the organisation of his work and that he persistently planned its publication, even publishing little in lifetime. This supports the hypothesis that his work does have a certain unity, revealing a consistent development and not a fragmentary purpose, as suggested by a large number of critics.

In the edition, this understanding is underpinned by the encoding of micro-genetic textual phenomena in a basic way to elucidate shifts in the conception of work titles, authorship and publication plans in single documents. Developments of the editorial projects on the macro level are traced by encoding mentioned entities in the documents and thereby establishing links between documents, publications and abstract works.

Different editorial procedures are combined to explore the potential of a digital edition which proves fruitful to bring the understanding and editorial presentation of Pessoa's work further than existing print editions, as well as contributing to clarify a major critical topic concerning the author's envisioned unity of his work.

Bibliography

Belknap, R. E. (2004). *The List. The Uses and Pleasures of Cataloguing.* New Haven and London: Yale University Press.

Castro, I. (2013). *Editar Pessoa.* 2nd Edition. Lisbon: Imprensa Nacional – Casa da Moeda.

Coelho, Jacinto do Prado (1949). *Diversidade e Unidade em Fernando Pessoa.* Lisbon: Ocidente.

Cunha, T. S. (1987). Planos e projectos editoriais de Fernando Pessoa: uma velha questão. *Revista da Biblioteca Nacional*, Series 2, **2** (1): 93-107.

Duarte, L. F. (1988). Texto acabado e texto virtual ou a cauda do cometa. *Revista da Biblioteca Nacional*, Series 2, **3** (3): 167-81.

Galhoz, M. A. (1993). A fortuna editorial pessoana e seus problemas. In Seabra, J. A. (Ed), *Mensagem. Poemas esotéricos.* Madrid, etc.: Coleção Arquivos, pp. 216-26.

Gusmão, M. (2003). O Fausto – um teatro em ruínas. *Românica*, **12**: 67-86.

Martins, F. C. (2003). Breves notas sobre a alta definição. *Românica*, **12**: 157-164.

Martins, F. C. (2011). Fernando Pessoa e o Original Perdido. *Tágides. Revista de Literatura, Cultura e Arte Portuguesas*, **1**: 89-100.

Robinson, P. (2013). The Concept of the Work in the Digital Age. *Ecdótica*, **10**: 13-41.

Sahle, P. (2013). *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.*Teil 2: Befunde, Theorie und Methodik. Norderstedt: Books on Demand.

Sepúlveda, P. (2013). *Os livros de Fernando Pessoa*. Lisbon: Ática.

KARREK: Building and Annotating a Kafka/Reference Corpus

J. Berenike Herrmann
bherrma1@gwdg.de
Göttingen University Germany, Germany

Gerhard Lauer
gerhard.lauer@phil.uni-goettingen.de
Göttingen University Germany, Germany

The proposed paper reports on philological and computational aspects of building and annotating a literary corpus. As part of the ongoing corpus-stylistic project Q-LIMO (Quantitative Analysis of Literary Modernity), the KARREK (Kafka/Referenzkorpus), a corpus that centers on German narrative Literature, is designed to enable comparative quantitative-stylistic analyses of Franz Kafka's prose. The corpus includes literary as well as non-literary texts and is tagged for part-of-speech (POS) and enriched by selected types of meta-data (e.g., author, title, date of publication, and [narrative] genre). The number of digital German literary text collections has been growing, and there are several repositories, such as the TextGrid Repository, the German Text Archive (DTA), as well as Gutenberg-DE and individual digital collections (e.g., the Central Catalogue of Digitized Prints [zvdd]). However, so far, no digital resource has been specifically tailored to the needs of quantitative Kafka-research. We offer such a resource, KARREK, which covers the entirety of Kafka's writings (including literary and non-literary ones) as well as a meaningful selection of Newer German Literature, putting a focus on Modernism (ca. 1880 – 1930). What is more, it caters to consistent and high-quality linguistic annotation, text-markup, and metadata. The corpus is hence a unique resource; it will be made publicly available.

Building and annotating the KARREK poses unique challenges typical for textual analysis in Digital Humanities: The first main task is a philological one, selecting the texts that allow meaningful and hypothesis-driven analyses of Franz Kafka's writings. Philological standards of corpus construction are especially high in terms of editorial detail and consideration of cultural, societal and philosophical contexts (cf., Engel and Auerochs, 2010) as well as linguistic ones (cf. Blahak, 2015). KARREK is hence balanced for factors such as canonicity, popularity, (narrative) genre, and linguistic variety. It strives for clarity in terms of literary edition. A special feature of the corpus is its representation of Kafka's reading habits: we assume that a writer's texts may reflect a dimension of 'constructive reading' ('produktive Rezeption', cf. Grimm, 1977), resulting in stylistic similarity with particular authors. For corpus compilation, we thus systematically review the available sources to determine which literary and non-literary works may be candidates for having influenced his writings (e.g., Blank 2001; Born 1990; Born et al., 1983), with ensuing data analysis to explore Kafka's position within a global network of texts and authors.

Our aim to form a reference corpus (with some degree of 'representativeness' for Newer German Literature, Prague Literature, Modern Literature, etc.) requires transforming a substantial amount of 'big literary data' into a consistent corpus. The philological questions transform into computational tasks where retro-digitization (and OCR) is concerned, as well as preprocessing of the data, with consistent metadata (available literature on metadata (e.g., Alemuh and Brett, 2015; Lazinger, 2001), as well as a uniform textual format, including parameters of textual markup (TEI-5). The entire corpus is going to be published in a flexible stand-off format (TCF) which allows further annotation.

For our means, the second main task is the reliable and accurate linguistic annotation for POS (with the STTS tag-set for German). Word class is a reliable indicator of register and genre variation (e.g., Biber and Conrad, 2009), with a high degree of accurate automatic annotation. Although there are high-quality POS-taggers available for German (e.g., Unigram, HMM, and Perceptron Taggers), most were trained on late 20th-Century news texts. To ensure accurate tagging of historically prior and literary texts, for training our taggers, we use as a first resource the POS-tagged corpus of the German Text Archive (DTA) which comprises historical and literary data. In addition, a round of manual error management (on a sample of ca. 40,000 words), as well as training of a CRF-tagger (MarMot) are administered to heighten reported accuracy of POS-tagging for our corpus. We are currently establishing a workflow comprising an eXist-database with an annotation interface, as well as a procedure for evaluation of tagger output for different sections of the corpus. Next to the compilation of the corpus, our project will thus offer more insight into POS-tagging of historical German narrative literary texts.

In DH, tackling representativeness in literary corpus building is an important research task, with theoretical background offered by corpus linguists and computational linguists (e.g., Biber, 1994; Evert, 2006; Lee, 2001). Although methods of 'big data' computing (cf. Anderson, 2008; Manovich, 2015) are extremely promising for digital literary studies, we suggest that in our field creating balanced and representative corpora is a vital question. Since there are only few best practice examples (for German, e.g. the DTA), and representativeness is a recurring issue that by no means may be considered 'once-and-for-all'-solved, the way in which our particular study tackles the issue may be of interest for many similar studies. In the prospective talk, we will thus address 'representativeness', reporting on the decisions made on text selection (author, titles, genre, canon/popularity, 'constructive reading').

We will also provide a description of the linguistic (POS) annotation, including details on our machine-learning procedure and the user interface. Also, the workflow will be presented, elucidating ways of applying it to analogous studies/projects. The talk will give a description of the online publication details. In all, our research shows that building and enriching a particular literary corpus is by far no trivial task, but requires a sound theoretical modeling of the phenomenon constructed and an interdisciplinary method that does justice to philological as well as to computational criteria of high-quality corpus research.

Bibliography

Alemuh, G., and Brett, S. (2015). *An Emergent Theory of Digital Library Metadata: Enrich the Filter*. Amsterdam: Elsevier.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 6 March 2016).

Biber, D. (1994). Representativeness in corpus design. In A. Zampolli, N. Calzolari, AND M. Palmer (Eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker*, pp. 377–407. Springer Netherlands. http://link.springer.com/chapter/10.1007/978-0-585-35958-8_20 (accessed 19 February 2016).

Biber, D., and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Blank, H. (2001). *In Kafkas Bibliothek: Werke der Weltliteratur und Geschichte in der Edition, wie sie Kafka besaß oder kannte; kommentiert mit Zitaten aus seinen Briefen und Tagebüchern*. Stuttgart: Blank.

Blahak, B. (2015). *Franz Kafkas Literatursprache: Deutsch im Kontext des Prager Multilingualismus*. Köln: Böhlau Verlag.

Born, J. (1990). *Kafkas Bibliothek: ein beschreibendes Verzeichnis; mit einem Index aller in Kafkas Schriften erwähnten Bücher, Zeitschriften und Zeitschriftenbeiträge*. Frankfurt am Main: S. Fischer.

Born, J., and Koch, E. (Eds.) (1983). *Franz Kafka: Kritik und Rezeption, 1924-1938*. Frankfurt am Main: S. Fischer.

Engel, M., and Auerochs, B. (Eds.) (2010). *Kafka-Handbuch. Leben, Werk, Wirkung*. Stuttgart: Metzler.

Evert, S. (2006). How Random is a Corpus? The Library Metaphor. *Zeitschrift für Anglistik und Amerikanistik*, **54**(2): 177–90.

Grimm, G. E. (1977). *Rezeptionsgeschichte: Grundlegung einer Theorie*. München: Fink.

Lazinger, S. S. (2001). *Digital Preservation and Metadata: History, Theory, Practice*. Englewood, Colo.: Libraries Unlimited.

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating through the BNC jungle. *Language Learning and Technology*, **5**(3): 37–72.

Manovich, L. (2015). Data science and digital art history. *International Journal for Digital Art History*, (1). http://nbn-resolving.de/urn:nbn:de:bsz:16-dah-216313 (accessed 6 March 2016).

Invisible Cities In Literature And History: Interfaces To Scalable Readings Of Textual And Visual Representations Of The Imaginary

Charles van den Heuvel
charles.van.den.heuvel@huygens.knaw.nl
Huygens Institute for the History of the Netherlands, Netherlands,

Florentina Armaselu
florentinaa@zoomimagine.com
Centre virtuel de la connaissance sur l' Europe (CVCE), Luxemburg

Introduction

At the DH2014 conference it was explained how the *ZoomImagine* software (Z-editor) allows zooming on units of text and contextual information (z-lexias) in order to explore layers of meaning that support critical readings of literature (Armaselu 2014). Changing perspective or analytical viewpoints by using different types of “magnifying glass” is also possible (Figure 1).

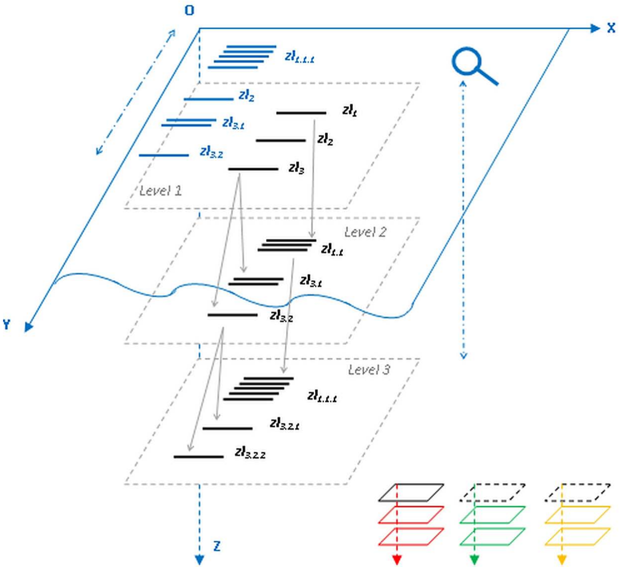
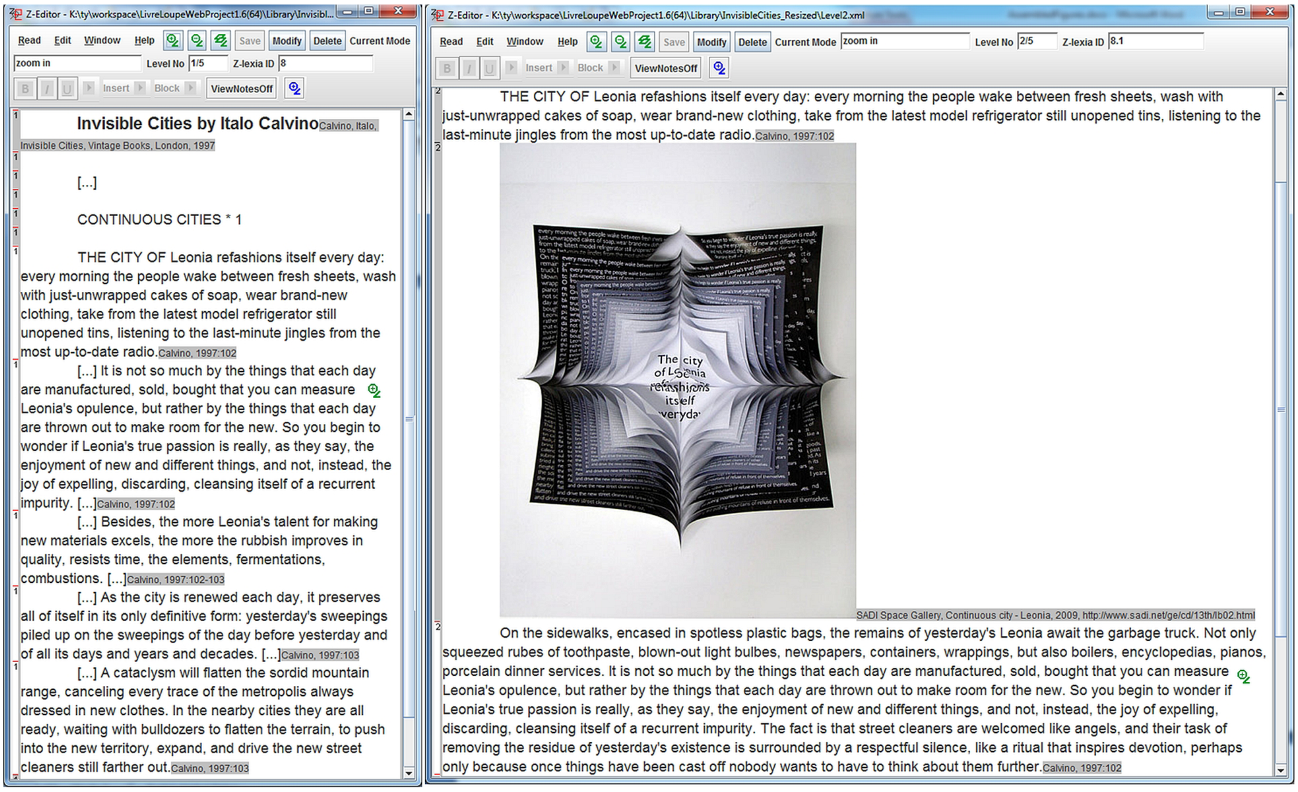


Figure 1. Z-text layout. Levels of z-lexias (center). Changing perspective (bottom right)

In this study the results of two experiments will be discussed in which the Z-editor was used to explore crossovers between text and image and history and literature in descriptions of imaginary cities and imaginary depictions of existing cities. The first case concerns readings of the *Invisible Cities* (1972) in which Italo Calvino (1923-1985) described Marco Polo’s accounts of visits to cities to Kubla Kahn, Emperor of the Tartars. The second one discusses designs for and historical depictions of the citadel of the city



Z-Editor - K:\ty\workspace\LivreLoupeWebProject1.6(64)\Library\Invisibl...
Read
Edit
Window
Help
Save
Modify
Delete
Insert
Block
ViewNotesOff

Current Mode zoom in Level No 1/5 Z-lexia ID 8

1 Invisible Cities by Italo Calvino Calvino, Italo, Invisible Cities, Vintage Books, London, 1997

1

1 [...]

1

1 CONTINUOUS CITIES * 1

1

1 THE CITY OF Leonia refashions itself every day: every morning the people wake between fresh sheets, wash with just-unwrapped cakes of soap, wear brand-new clothing, take from the latest model refrigerator still unopened tins, listening to the last-minute jingles from the most up-to-date radio.Calvino, 1997:102

1 [...] It is not so much by the things that each day are manufactured, sold, bought that you can measure Leonia's opulence, but rather by the things that each day are thrown out to make room for the new. So you begin to wonder if Leonia's true passion is really, as they say, the enjoyment of new and different things, and not, instead, the joy of expelling, discarding, cleansing itself of a recurrent impurity. [...]Calvino, 1997:102

1 [...] Besides, the more Leonia's talent for making new materials excels, the more the rubbish improves in quality, resists time, the elements, fermentations, combustions. [...]Calvino, 1997:102-103

1 [...] As the city is renewed each day, it preserves all of itself in its only definitive form: yesterday's sweepings piled up on the sweepings of the day before yesterday and of all its days and years and decades. [...]Calvino, 1997:103

1 [...] A cataclysm will flatten the sordid mountain range, canceling every trace of the metropolis always dressed in new clothes. In the nearby cities they are all ready, waiting with bulldozers to flatten the terrain, to push into the new territory, expand, and drive the new street cleaners still farther out.Calvino, 1997:103

Z-Editor - K:\ty\workspace\LivreLoupeWebProject1.6(64)\Library\InvisibleCities_Resized\Level2.xml
Read
Edit
Window
Help
Save
Modify
Delete
Insert
Block
ViewNotesOff

Current Mode zoom in Level No 2/5 Z-lexia ID 8.1

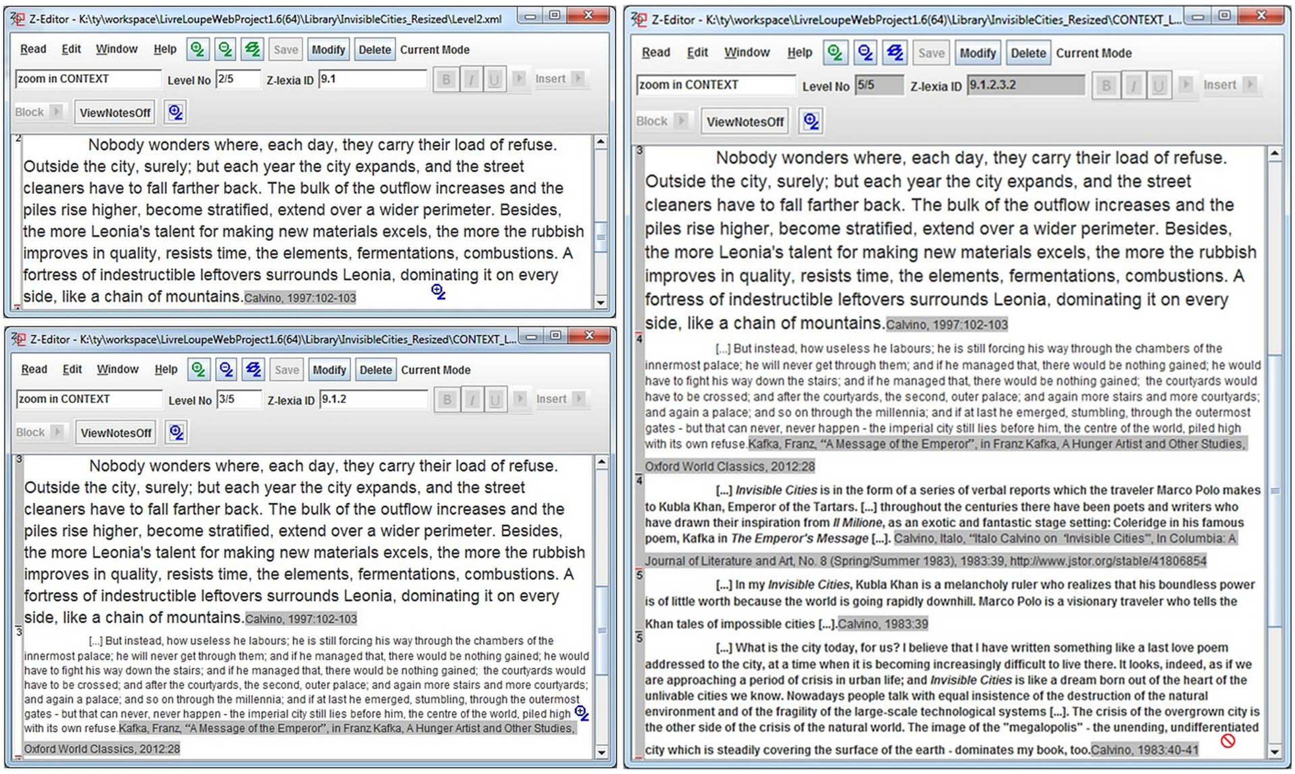
2 THE CITY OF Leonia refashions itself every day: every morning the people wake between fresh sheets, wash with just-unwrapped cakes of soap, wear brand-new clothing, take from the latest model refrigerator still unopened tins, listening to the last-minute jingles from the most up-to-date radio.Calvino, 1997:102

2

SADI Space Gallery, Continuous city - Leonia, 2009, http://www.sadi.net/ge/cd/13th/lb02.html

2 On the sidewalks, encased in spotless plastic bags, the remains of yesterday's Leonia await the garbage truck. Not only squeezed rubes of toothpaste, blown-out light bulbes, newspapers, containers, wrappings, but also boilers, encyclopedias, pianos, porcelain dinner services. It is not so much by the things that each day are manufactured, sold, bought that you can measure Leonia's opulence, but rather by the things that each day are thrown out to make room for the new. So you begin to wonder if Leonia's true passion is really, as they say, the enjoyment of new and different things, and not, instead, the joy of expelling, discarding, cleansing itself of a recurrent impurity. The fact is that street cleaners are welcomed like angels, and their task of removing the residue of yesterday's existence is surrounded by a respectful silence, like a ritual that inspires devotion, perhaps only because once things have been cast off nobody wants to have to think about them further.Calvino, 1997:102

Figure 2. *Invisible Cities* Z-text. Level 1 (left), 2 (right). Calvino's text unfolding along different lines of "stratification", the "new-old" divide and the "textual-visual" metamorphosis (the visual "becoming" of the text). The first level (left) contains fragments (z-lexias) conveying the meaning of "new" (through words like "new", "renew"). By zooming-in (right), further elements from deeper levels can be surfaced, i.e. fragments entailing the idea of accumulation of "old" (including words such as "yesterday", "past", "residue", "remains", "garbage", "refuse") or visual representations not belonging to the original text but imagined by various artists.



Z-Editor - K:\ty\workspace\LivreLoupeWebProject1.6(64)\Library\InvisibleCities_Resized\Level2.xml
Read
Edit
Window
Help
Save
Modify
Delete
Insert
Block
ViewNotesOff

Current Mode zoom in CONTEXT Level No 2/5 Z-lexia ID 9.1

2 Nobody wonders where, each day, they carry their load of refuse. Outside the city, surely; but each year the city expands, and the street cleaners have to fall farther back. The bulk of the outflow increases and the piles rise higher, become stratified, extend over a wider perimeter. Besides, the more Leonia's talent for making new materials excels, the more the rubbish improves in quality, resists time, the elements, fermentations, combustions. A fortress of indestructible leftovers surrounds Leonia, dominating it on every side, like a chain of mountains.Calvino, 1997:102-103

Z-Editor - K:\ty\workspace\LivreLoupeWebProject1.6(64)\Library\InvisibleCities_Resized\CONTEXT_L...
Read
Edit
Window
Help
Save
Modify
Delete
Insert
Block
ViewNotesOff

Current Mode zoom in CONTEXT Level No 3/5 Z-lexia ID 9.1.2

3 Nobody wonders where, each day, they carry their load of refuse. Outside the city, surely; but each year the city expands, and the street cleaners have to fall farther back. The bulk of the outflow increases and the piles rise higher, become stratified, extend over a wider perimeter. Besides, the more Leonia's talent for making new materials excels, the more the rubbish improves in quality, resists time, the elements, fermentations, combustions. A fortress of indestructible leftovers surrounds Leonia, dominating it on every side, like a chain of mountains.Calvino, 1997:102-103

3 [...] But instead, how useless he labours; he is still forcing his way through the chambers of the innermost palace; he will never get through them; and if he managed that, there would be nothing gained; he would have to fight his way down the stairs; and if he managed that, there would be nothing gained; the courtyards would have to be crossed; and after the courtyards, the second, outer palace; and again more stairs and more courtyards; and again a palace; and so on through the millennia; and if at last he emerged, stumbling, through the outermost gates - but that can never, never happen - the imperial city still lies before him, the centre of the world, piled high with its own refuse.Kafka, Franz, "A Message of the Emperor", in Franz Kafka, A Hunger Artist and Other Studies, Oxford World Classics, 2012:28

Z-Editor - K:\ty\workspace\LivreLoupeWebProject1.6(64)\Library\InvisibleCities_Resized\CONTEXT_L...
Read
Edit
Window
Help
Save
Modify
Delete
Insert
Block
ViewNotesOff

Current Mode zoom in CONTEXT Level No 5/5 Z-lexia ID 9.1.2.3.2

3 Nobody wonders where, each day, they carry their load of refuse. Outside the city, surely; but each year the city expands, and the street cleaners have to fall farther back. The bulk of the outflow increases and the piles rise higher, become stratified, extend over a wider perimeter. Besides, the more Leonia's talent for making new materials excels, the more the rubbish improves in quality, resists time, the elements, fermentations, combustions. A fortress of indestructible leftovers surrounds Leonia, dominating it on every side, like a chain of mountains.Calvino, 1997:102-103

4 [...] But instead, how useless he labours; he is still forcing his way through the chambers of the innermost palace; he will never get through them; and if he managed that, there would be nothing gained; he would have to fight his way down the stairs; and if he managed that, there would be nothing gained; the courtyards would have to be crossed; and after the courtyards, the second, outer palace; and again more stairs and more courtyards; and again a palace; and so on through the millennia; and if at last he emerged, stumbling, through the outermost gates - but that can never, never happen - the imperial city still lies before him, the centre of the world, piled high with its own refuse.Kafka, Franz, "A Message of the Emperor", in Franz Kafka, A Hunger Artist and Other Studies, Oxford World Classics, 2012:28

4 [...] *Invisible Cities* is in the form of a series of verbal reports which the traveler Marco Polo makes to Kubla Khan, Emperor of the Tartars. [...] throughout the centuries there have been poets and writers who have drawn their inspiration from *Il Milione*, as an exotic and fantastic stage setting: Coleridge in his famous poem, Kafka in *The Emperor's Message* [...]. Calvino, Italo, "Italo Calvino on 'Invisible Cities'", In Columbia: A Journal of Literature and Art, No. 8 (Spring/Summer 1983), 1983:39, http://www.jstor.org/stable/41806854

5 [...] In my *Invisible Cities*, Kubla Khan is a melancholy ruler who realizes that his boundless power is of little worth because the world is going rapidly downhill. Marco Polo is a visionary traveler who tells the Khan tales of impossible cities [...].Calvino, 1983:39

5 [...] What is the city today, for us? I believe that I have written something like a last love poem addressed to the city, at a time when it is becoming increasingly difficult to live there. It looks, indeed, as if we are approaching a period of crisis in urban life; and *Invisible Cities* is like a dream born out of the heart of the unlivable cities we know. Nowadays people talk with equal insistence of the destruction of the natural environment and of the fragility of the large-scale technological systems [...]. The crisis of the overgrown city is the other side of the crisis of the natural world. The image of the "megalopolis" - the unending, undifferentiated city which is steadily covering the surface of the earth - dominates my book, too.Calvino, 1983:40-41

Figure 3. *Invisible Cities* Z-text. Level 2 (top left), 3 (bottom left), 5 (right). "Multidimensional" exploration by switching perspectives. One can zoom-in to get the visual representation of a fragment (compare Figure 2) or, instead, change the perspective and zoom-in with a contextual magnifying glass and expand Calvino's text with Kafka's or Calvino's own comments on *Invisible Cities* (Kafka, 2012; Calvino, 1983).

of Groningen in the Netherlands in an atlas of the Flemish engineer Pierre Lepoivre (1546-1626). We will demonstrate that the Z-text model is both suitable for analyzing critical readings and artistic impressions of the Invisible Cities as for assessing different levels of historical evidence of visualizations, such as maps, drawings and designs of existing cities. Furthermore, it allows for navigating in an associative way through imaginary cities of which features might have inspired both Calvino's literary writings as Lepoivre's designs. Contemporary artists still inspired by Calvino seem to delve into a collective visual memory of "cities", in a way Renaissance engineers linked their designs to representations of ideal cities and new fortress towns, from Campanella's City of the Sun to Zamość in Poland. The Z-text model might stimulate to explore digital past and future identities of imaginary and real cities. We will conclude with a discussion of the potential of the Z-editor interface for general debates within the digital humanities.

The Invisible Cities of Italo Calvino

The publication of *Le Città Invisibili* in 1972 by Italo Calvino was directly followed by comments by literary critics and by visualizations of artists inspired by the poetical descriptions of the imaginary cities in the imageless book. The division into 11 themes associated with the city that return each 5 times in the book -City and Memory, City and Desire, City and Signs, Thin Cities, Trading Cities, Cities and Eyes, Cities and Names, Cities and the Dead, Cities and the Sky, Continuous Cities and Hidden Cities- has resulted into various interpretations of these imaginary urban spaces, such as "rhizomatic space" Kerstin Pilz (2003) and "city of strata" Sambit Panigrahi (2014). The zooming functionalities of the editor and Z-text model make it possible for instance to analyze Calvino's description of the continuous city of Leonia that each days breaches through its boundaries of deposited waste at its circumference with these contemporary interpretations (Figure 2).

Another "rhizomatic" expansion imagined via the model - along with the "contextualization" line - enables other readings of Calvino's work, both by critics and himself. Calvino in some occasions looks through the lens of the writer and in other moments through the eyes of the reader or editor and states "that the author's view no longer counts" (Baldi, 2015; Calvino, 2015: 41-42). Calvino sometimes puts the Invisible Cities directly into context by referring to works that similarly are inspired by Marco Polo's travels, such as the poem *Kubla Kahn* (1798) of Samuel Taylor Coleridge or *The Message from the Emperor* (1919) by Franz Kafka. In other occasions references are indirect, for example when Calvino notes that the atlas of Kubla Kahn contains images of "lands visited in thought, but not yet discovered or founded: New Atlantis, Utopia, the City of the Sun [..]" (Calvino, 1997: 147). On a more detailed level the possibility of sideways movements becomes important when Calvino reacts to all those critics who underlined the importance of the closing sentence by claiming that the Invisible Cities is "a many facetted book" with "various possible 'conclusions'" (Calvino, 1983: 41). By combining the zooming functionality with the representation of contextual information a Z-text becomes a multidimensional space of analysis (Figure 3).

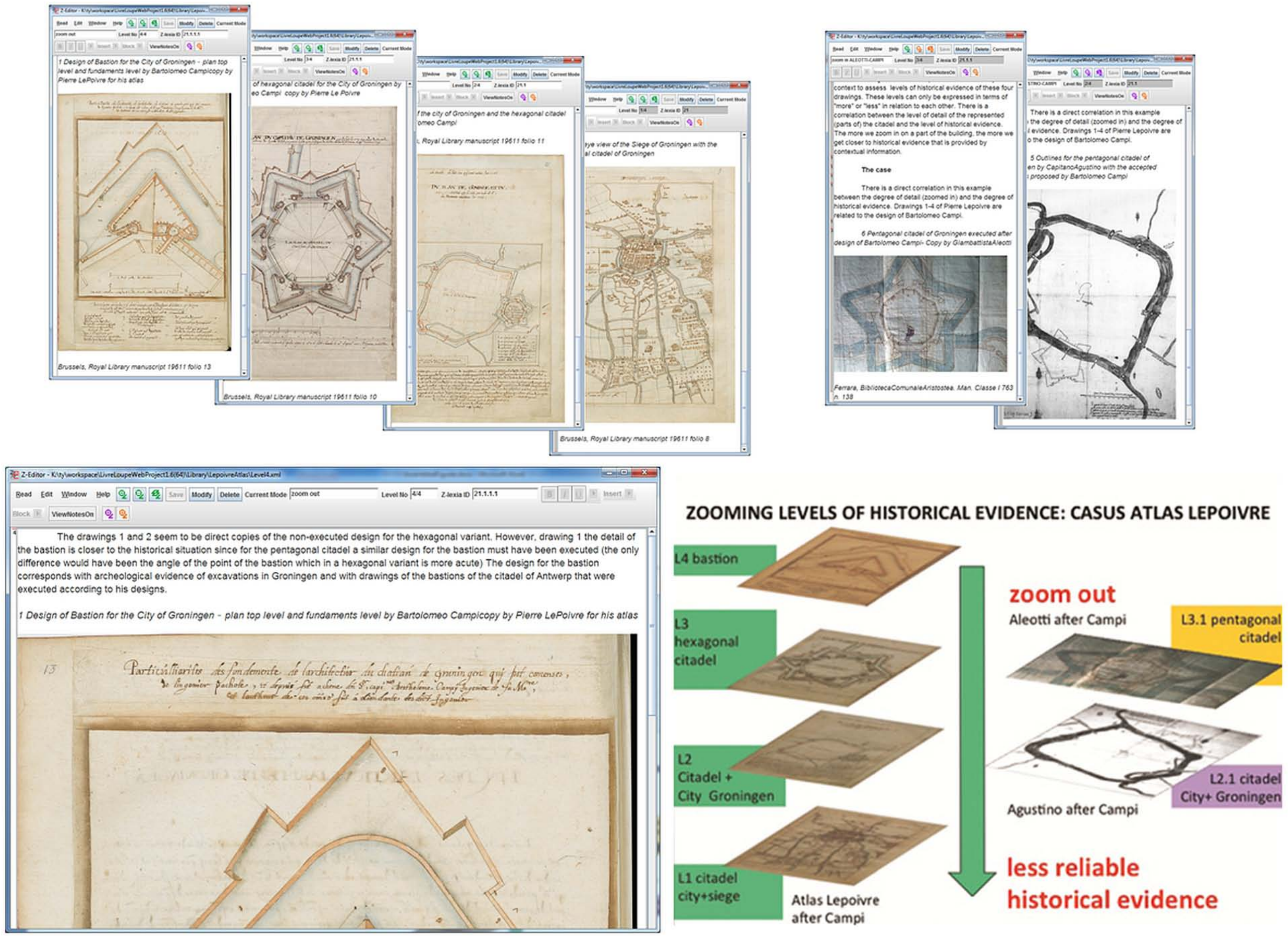
Such an interface may be useful for the analysis of the Invisible Cities when Calvino observes: "And yet, all these pages put together did not make a book: for a book (I think) is something which has a beginning and an end [..] It is a space which the reader must enter, wander round, maybe lose his way in, and eventually find an exit, or perhaps even several exits, [..]" (Calvino, 1983: 38). If we enter our concentric city of Leonia again, new associative perspectives pop up. We recognize Calvino's reference to Kafka's story of the Emperor's messenger wanting to report about the death of Kubla Kahn by trying - breaking through the walls surrounding the palace, "the center of the world, piled high with its own refuse." (Kafka, 2012: 28) However, another of Calvino's references, the one to Campanella's concentric walled City of the Sun allows to link the form of Leonia with images of built and not built fortified cities that are related to our second experiment dealing with levels of historical evidence of designs of citadels in the atlas of Lepoivre (compare Figure 5). This comparison is of interest since Calvino once considered a theme "Cities and Form", but decided to merge it with those of other cities (Calvino, 1983: 38).

Imaginary Depictions of An Existing City: Groningen (Netherlands)

Renaissance ideal-fortress cities were ideally radial-concentric cities with a polygonal perimeter with angular bastions. Therefore circumscribing and inscribing circles played an essential role in designing fortified cities and citadels. The design of fortifications was not just functional. Fortification atlases were cultural artifacts amongst other books on architecture such as the Renaissance Vitruvius editions that often contained visual interpretations of the city described in the lost original manuscript. In this context also the four drawings based on the designs of the Italian engineer Bartolomeo Campi of the city and citadel of Groningen in the Atlas of Lepoivre of ca. 1624 must be seen (Heuvel, 1994: 1998) (Figure 4).

The Z-text model and the digital humanities

The Z-text model meets a wide array of new reading and analysis practices discussed in relation to the digital humanities. While the zooming function supports bridging distant and close reading by scalable reading (Mueller, 2012) the combination with contextualization on the various planes to read text and image from various perspectives coincides with notions of deep reading



Z-Editor - K:\ty\workspace\LivreLoupeWebProject1.6(64)\Library\LepoivreAtlas\Level4.xml

Read Edit Window Help Save Modify Delete Current Mode zoom out Level No 4/4 Z-lexia ID 21.1.1.1 Insert Block ViewNotesOn

The drawings 1 and 2 seem to be direct copies of the non-executed design for the hexagonal variant. However, drawing 1 the detail of the bastion is closer to the historical situation since for the pentagonal citadel a similar design for the bastion must have been executed (the only difference would have been the angle of the point of the bastion which in a hexagonal variant is more acute) The design for the bastion corresponds with archeological evidence of excavations in Groningen and with drawings of the bastions of the citadel of Antwerp that were executed according to his designs.

1 Design of Bastion for the City of Groningen - plan top level and fundaments level by Bartolomeo Campicopy by Pierre LePoivre for his atlas

Brussels, Royal Library manuscript 19611 folio 13

of hexagonal citadel for the City of Groningen by eo Campi copy by Pierre Le Poivre

Brussels, Royal Library manuscript 19611 folio 10

f the city of Groningen and the hexagonal citadel omeo Campi

, Royal Library manuscript 19611 folio 11

eye view of the Siege of Groningen with the al citadel of Groningen

Brussels, Royal Library manuscript 19611 folio 8

context to assess levels of historical evidence of these four drawings. These levels can only be expressed in terms of "more" or "less" in relation to each other. There is a correlation between the level of detail of the represented (parts of) the citadel and the level of historical evidence. The more we zoom in on a part of the building, the more we get closer to historical evidence that is provided by contextual information.

The case

There is a direct correlation in this example between the degree of detail (zoomed in) and the degree of historical evidence. Drawings 1-4 of Pierre Lepoivre are related to the design of Bartolomeo Campi.

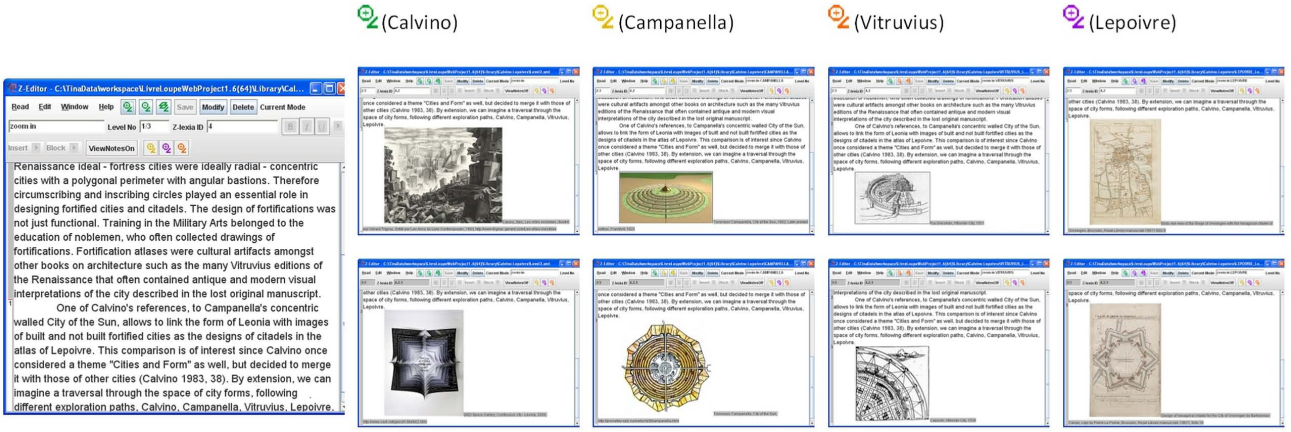
6 Pentagonal citadel of Groningen executed after design of Bartolomeo Campi- Copy by GiambattistaAleotti

Ferrara, BibliotecaComunaleAristotea, Man. Classe I 763 n. 138

5 Outlines for the pentagonal citadel of en by CapitanoAgustino with the accepted proposed by Bartolomeo Campi

ZOOMING LEVELS OF HISTORICAL EVIDENCE: CASUS ATLAS LEPOIVRE
L4 bastion
L3 hexagonal citadel
L2 Citadel + City Groningen
L1 citadel city+siege
Atlas Lepoivre after Campi
zoom out
Aleotti after Campi
L3.1 pentagonal citadel
L2.1 citadel City+ Groningen
Agustino after Campi
less reliable historical evidence

Figure 4. *Lepoivre Atlas.* Z-text layout (top), level 4 (bottom left); interpretation (bottom right). The imaginary space of historical representation can be traversed through layers corresponding to decreasing levels of historical evidence (top left) or to different perspectives (Campi/Aleotti and Agustino variants, top right).



Z-Editor - C:\TinaData\workspace\LivreLoupeWebProject1.6(64)\Library\Cal...

Read Edit Window Help Save Modify Delete Current Mode zoom in Level No 1/3 Z-lexia ID 4 Insert Block ViewNotesOn

Renaissance ideal - fortress cities were ideally radial - concentric cities with a polygonal perimeter with angular bastions. Therefore circumscribing and inscribing circles played an essential role in designing fortified cities and citadels. The design of fortifications was not just functional. Training in the Military Arts belonged to the education of noblemen, who often collected drawings of fortifications. Fortification atlases were cultural artifacts amongst other books on architecture such as the many Vitruvius editions of the Renaissance that often contained antique and modern visual interpretations of the city described in the lost original manuscript.

One of Calvino's references, to Campanella's concentric walled City of the Sun, allows to link the form of Leonia with images of built and not built fortified cities as the designs of citadels in the atlas of Lepoivre. This comparison is of interest since Calvino once considered a theme "Cities and Form" as well, but decided to merge it with those of other cities (Calvino 1983, 38). By extension, we can imagine a traversal through the space of city forms, following different exploration paths, Calvino, Campanella, Vitruvius, Lepoivre.

Figure 5. *Imaginary Cities* Z-text. Level 1 (left), 2 (top right), 3 (bottom right). Traversal through the space of forms, following multiple exploration paths and connecting imaginary cities in literature (Calvino, Campanella) and in architectural treatises and theory (Vitruvius) with imaginary representations of existing cities in design (Lepoivre).

(Birkets, 1994). This combination allows for the creation of multiple levels of meaning and supports a continuously process of reinterpretation from multiple perspectives, contributing this way in developing digital hermeneutic methods (Capurro, 2010).

Bibliography

Armaselu, F. (2014). The Layered Text. From Textual Zoom, Text Network Analysis and Text Summarisation to a Layered Interpretation of Meaning. *Digital Humanities 2014 Conference Abstract EPFL-UNIL*, Lausanne, Switzerland, pp. 79-82. http://dharchive.org/paper/DH2014/Paper-515.xml. (accessed 2 March 2016)

Baldi, E. (2015). lL’ ochio che scrive. La dinamica dell’ imagine autoriale di Calvino nella critica italiana, *Incontri* 2015, **1**: 22-33www.rivista-incontri.nl (accessed 2 March 2016)

Birkerts, S. (1994). *The Gutenberg Elegies: the Fate of Reading in an Electronic Age*, Faber and Faber, Boston (MA).

Capurro, R. (2010). Digital Hermeneutics: An outline. *AI and Society*, 2010, **35**(1): 35-42, http://www.capurro.de/digital-hermeneutics.html.

Calvino, I. (1972). *Le città invisibili*, Einaudi, Turin.

Calvino, I. (1983). Italo Calvino on ‘Invisible Cities’. *Columbia: A Journal of Literature and Art*, **8**: 7-42.
http://www.jstor.org/stable/41806854 (accessed 2 March 2016).

Calvino, I. (1997). *Invisible Cities. Translated from the Italian by William Weaver*. London: Vintage Books.

Heuvel, Charles van den (1994). Bartolomeo Campi successor to Francesco Paciotto. A different method of designing citadels: Groningen and Flushing. In: *Architetti e ingegneri militari italiani all'estero dal XV al XVIII secolo*. Livorno-Roma, pp. 153-67.

Heuvel, Charles van den (1998). Pierre Le Poivre (1546-1626). Engineer of the King and the Representation of Architecture, In W. Thomas, L. Duerloo, (Eds.), *Albrecht and Isabella. Essays*. Turnhout, pp. 198-202.

Kafka, F. (2012). A Message of the Emperor. In Franz Kafka, *A Hunger Artist and Other Studies. A new translation by Joyce Crick*, World Classics 2012, 28. (First published 1919).

Lepoivre, P. (1624). *Les Plans des Villes des Païs de Hennault, d’ Artois, de Breband, très noblemen descripts a la plume par l’ architecte Pierre Le Poivre* etc. Brussels, Royal Library Albert I, Ms. 19611.

Mueller, M. (2012). Scalable Reading. Dedicated to Data, digitally assisted text analysis, https://scalablereading.northwestern.edu/scalable-reading/ (accessed 2 March 2016).

Panigrahi, S. (2014). Cities as Strata in Italo Calvino's Invisible Cities. *The Explicator*, **72**(1): 23-7, doi: 10.1080/00144940.2013.875873.

Pilz, K. (2003). Reconceptualising Thought and Space: Labyrinths and Cities in Calvino's Fictions. *Italica*, **80**: 229-42, http://www.jstor.org/stable/30038769 (accessed 2 March 2016).

ZoomImagine: www.zoomimagine.com (accessed 2 March 2016).

Silva Portentosissima – Computer-Assisted Reflections on Bifurcativity in Stemmas

Armin Hoenen
hoenen@em.uni-frankfurt.de
Goethe Universität Frankfurt, Germany

Introduction

In 1928, the philologue Joseph Bédier explored contemporary stemmas and found them to contain a suspiciously large amount of bifurcations. He gave two potential reasons, one being that the editors constructing a stemma with a bifurcation below the root node, could freely choose the variants for their base text. The second reason termed *force dichotomique* postulates that constructing a stemma, because editors are always comparing manuscript pairs, they tend to overseparate. This is illustrated in Figure 1. The philologue would postulate a common ancestor for the closer pair and attach the third sibling together with this ancestor to the parent node producing two bifurcations, where there was a multifurcation originally.

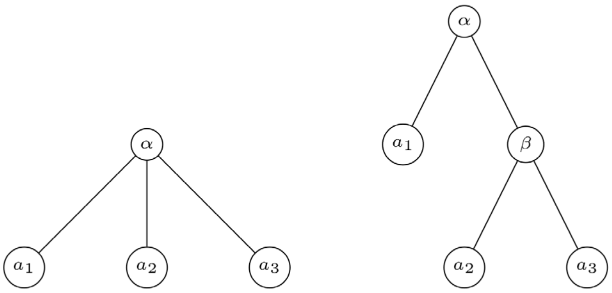


Figure 1. Force dichotomique, example. Left: true stemma. Right: probable philological reconstruction

(Maas, 1937) acknowledges a large amount of bifurcations in stemmas for recensions of Greek texts, but points out that this is rather unsurprising taking into account the number of possible stemmas and the proportion of bifurcations therein. (Felsenstein, 1978) computes the amounts of entirely bifurcating trees for *n* labelled leafs, which is certainly correlated with the number of predominantly bifurcating trees (and of root-bifurcating trees). As one can see, the proportion of bifurcating trees steadily declines, weakening Maas’ argument.

(Haugen, 2015) analysed collections of stemmas and found that Bédiers reason number two seems to hold independent of provenience. (Trovato, 2014) found that a realistic estimate of the percentage of lost manuscripts for medieval and earlier traditions ranges realistically above 73%. Related discussions and recent developments are reflected in projects such as Stemmaweb and the Parvum lexicon Stemmatologicum.

In this paper, the argument is introduced and investigated that, with a large amount of lost manuscripts, the amount of bifurcations in the true stemmas could naturally be high because of the transformations stemmas undergo when manuscript loss prunes away whole branches and leafs at a rate higher than 73%.

Distributions of outdegrees of manuscripts

Two basic distributions are most obviously historically interpretable, a normal distribution and an exponential distribution. In considering the outdegrees to be normally distributed, we assume, that there is one certain number of copies which is most probable, peaking the others; the more the outdegree diverges negatively or positively, the fewer manuscripts with this outdegree will be found. This could translate into a hypothetical historical projection, where many manuscripts of a tradition were copied and wore off at similar rates. Assuming an exponential distribution, things become more hierarchical. For instance, a powerful organization declares some of the manuscripts authoritative which would lead to certain manuscripts being copied many times more than others.

(Haugen, 2015) lists a table with the numbers of furcations (though not leafs) in two collections of stemmas for editions of Old Norse texts. Transforming these furcations into vectors, we conduct the Shapiro-Wilk normality test, (Shapiro and Wilk, 1965), and the exponentiality test by Kolmogorov-Smirnov (which are both robust even for smaller sample sizes) using the software R.[1] This results in an estimate on which distribution these stemmas entail for manuscript copying. However, stemmas with only identical values or only 2 furcations are skipped due to not being testable with the above chosen tests. Table 1 reveals a tendency towards exponentiality. Applying other tests for additional distributions, the weibull and log-normal distributions fitted the data best in terms of likelihood, suggesting a similar scenario.[2]

Value	Bibliotheca Arnamagnæna series	Editiones Arnamagnæna series
Normality	2	2
Exponentiality	11	9
Both	7	7
None	14	15
Stemmas	34	33

Table 1: Tests of distributions of furcations in the collections investigated by (Haugen, 2015). Tests at significance level 0.05

Experiment

In order to assess the above mentioned hypothesis, we need large numbers of underlying stemmas from which loss can be simulated and an appropriate model of manuscript loss. Then, we can simulate a large number of stemmas and manuscript loss, whereafter we count the number of bifurcations. Using simulations especially in large scale scenarios is sometimes the only possibility of assessing historical data, especially considering limits imposed onto empirical studies through data sparsity and other hindances.

Manuscript loss

Historical loss does not affect all manuscipts evenly. (Canfora, 2002) found private exemplars to be less affected whereas since libraries tend to be burnt in wars, public exemplars suffer loss more easily. Many other factors (climate, authoritativity, etc.) excert influence on manuscript loss most of which have never been made subject to generalizable quantifications and it is questionable if this can ever be done. Herein, we elaborate a simple model of loss using only two basic assumptions. We simulate loss of 73-100% where each node gets a probability related to its age (the older the more probably lost) and its outdegree (the more copied, the more probably kept in good conditions, the less probably lost). Since aging is considered to be stronger than preservational effort, we square the age dependent parameter. The probability of loss for each node is thus determined by

$$l_i^2 * slow(i)$$

where l is the height of the current node i incremented by 1 and slow(i) is the outdegree slowdown function:

$$slow(i) = \begin{cases} \frac{1}{o(i)} & \text{if } o(i) > 0 \\ 1 & \text{else} \end{cases}$$

where o(i) is the outdegree of node i. Note, that there is no distinction for nodes with an outdegree of one and leafs. This model of loss produced rather desirable loss probability distributions as is exemplified in the Appendix. However, we also use pure randomization of loss with equal probabilities among all nodes and randomization for probabilities as described but without squaring.

Simulation

For the simulation, we use R and Java.[3] First, for simulating normally distributed copying, we generate distributions randmonly drawn from a normal distribution using the R function *rnorm*. Since *rnorm* produces real numbers and negative numbers, we round all values and leave negative values and zeros aside. Since this may lead to a distortion of the so-sampled distribution deviating from a normal distribution, we test for the desired shape and only keep distributions which have a p-value above 0.05. This distribution is now our distribution of outdegrees in the to be simulated stemma, each value represents one outdegree. Starting with root, we randomly choose an outdegree (or make the node (not root) a leaf with a probability of 0.2) and add as many children to the actual node as this outdegree. For each of the so-generated children, we draw another outdegree and add as many children, recording them in the next generation. We iterate the process until all outdegrees are applied exactly once.This results in a differing number of leafs and a differing size of the tradition for each simulation. Since medieval traditions were probably not equal in size, the effect of this sampling is not controlled for further.[4]

With each tradition, loss is simulated in the three above described ways. We keep all nodes, which are on the path from root to any survivor but delete all other lost nodes. What remains is the true reconstructible stemma (TRS). Since philologues do not have sufficient information (and probably time) for the reconstruction of entirely lost branches, which would be too spurious an endeavour anyway, the stemma we simulated is the maximally faithful reconstruction given our simulated ground truth.

Distribution	Loss A	Loss B	Loss C
Normal (n=5, m=2, sd=1)	0.2 0.27	0.21 0.22	0.18 0.17
Normal (n=5, m=4, sd=1)	0.28 0.37	0.29 0.32	0.27 0.26
Exponential (n=5, r=0.5)	0.18 0.19	0.2 0.2	0.18 0.16
Exponential (n=5, r=1)	0.08 0.09	0.11 0.11	0.08 0.07

Table 2: Percentages of bifurcations and root-bifurcations given the parametricized simulations

Distribution	1	2	3	4	5	6
Normal (n=5, m=2, sd=1)	1216 2158	1740 546	1181 39	406 2	42 0	1 0
Normal (n=5, m=4, sd=1)	34 1852	359 881	1201 308	1757 97	1205 2	409 0
Exponential (n=5, r=0.5)	1509 1875	911 441	589 103	361 16	220 8	129 1
Exponential (n=5, r=1)	2052 1879	892 166	333 11	120 0	46 0	14 0

Table3: Furcations (1-6) before and after loss, Loss A

Results and discussion

Results can be seen in Tables 2 and 3. The simulated traditions ranged in size from 4 to more than 60 and after loss from 1 to around 20 manuscripts. Percentages of bifurations (even excluding leafs, as in Haugen (2015)) are below the observed values for the collections examined in the previous literature. Under more keeping nodes, which are on the path to root for any survivor as reconstructibles results in a much larger incidence of unifurcations, not bifurcations in the stemmas after loss has applied. Therefore, some model of contraction shall be applied in subsequent research. With a mean of 2 producing large numbers of bifurcations, after loss, their number is reduced sharply in contrast with unifurcations and leafs. Thus, *loss at high rates generally affects furcations of higher orders much more, leading to a naturally higher incidence of bi- and unifurcations and most of all obscuring an underlying distribution of original copies.* Aditionally, probably there are too few reconstructions of nodes with indegree and outdegree 1. In other words, philologues could tend to view a copy rather as sloppy than as a copies copy, which would contribute to making bifurcations more common in actual stemmas.

Conclusion

Although the simulation is an approximation, the values suggest, that indeed the proportion of bifurcations is suspiciously large in the collections in the literature. On the other hand, Maas was probably right in expecting a larger amount of lower order furcations. However, not their proportion among all possible stemmas, but the effect of massive loss of manuscripts leads to a large probable proportion of those in a pruned TRS.

Appendix

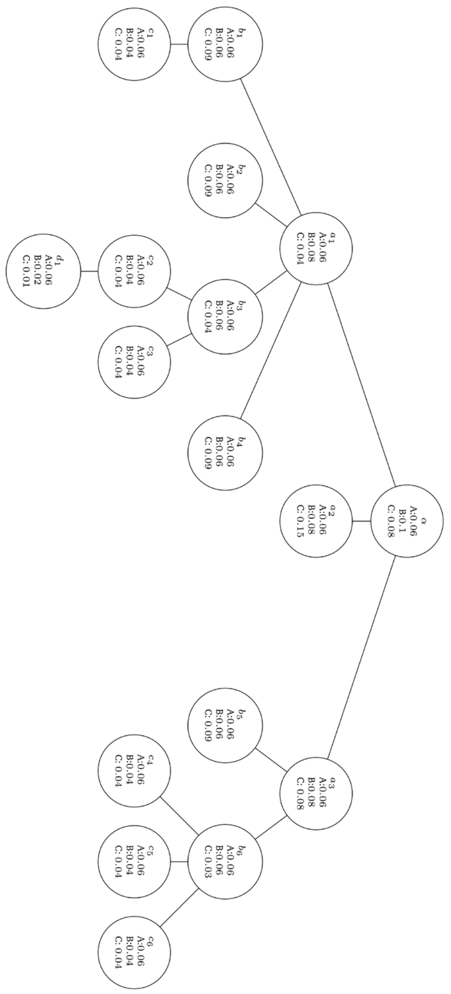


Figure 2. Simulation of manuscript loss. Probabilities of loss, according to models. Model A: randomized. Model B: age dependent. Model C: age dependent but slowed down by outdegree.

Bibliography

Bédier, J. (1928). La tradition manuscrite du 'Lai de l'Ombre': Réflexions sur l'Art d' Éditer les Anciens Textes. *Romania*, 394: 161-96, 321-56.

Canfora, L. (2002). *Il copista come autore*. Palermo: Sellerio.

Felsenstein, J. (1978). The number of evolutionary trees. *Systematic Zoology*, **27**(1): 27–33.

Haugen, O. E. (2015). The silva portentosa of stemmatology bifurcation in the recension of old norse manuscripts. *Digital Scholarship in the Humanities*, **30**(2).

Maas, P. (1937). Leitfehler und Stemmatische Typen. *Byzantinische Zeitschrift*, **37**(2): 289–94.

Satman, M. H. (2014). Rcaller: A software library for calling r from java. *British Journal of Mathematics & Computer Science*, **4**(15): 2188–96.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**: 591–611.

Trovato, P. (2014). *Everything You Always Wanted to Know about Lachmann's Method, A Non-Standard Handbook of Genealogical Textual Criticism in the Age of Post-Structuralism, Cladistics, and Copy-Text*. libreriauniversitaria.it.

Notes

[1] Packages exptest and fitdistr have been used.

[2] The average mean of the furcation distributions was 1.95 and the average standard deviation 0.78.

[3] We used JgraphT and RCaller, (Satman, 2014).

[4] One could compare the proportion of surviving to lost nodes in the TRS and stemmas from the literature.

Live/Life Stories. The Uses Of Digital War Testimonies In Educational Contexts

Susan Hogervorst
susan.hogervorst@ou.nl
Open Universiteit Nederland, Erasmus University Rotterdam, Netherlands, The

Since the 1980s, under pressure of the definite 'disappearance' of the eye witness generations of WWII, a wide variety of initiatives has been undertaken to capture eye witness memories for the future. Multiple oral history collections have been created throughout the western world, in which ten thousands of interviews and life stories have been preserved on audio and video. Parallel to the quest for individual war memories to collect and preserve, there has been an increasing effort to transmit these memories onto younger generations. In my dissertation, I have referred to this process as the pedagogization of memory, with which I point at the transmission of WWII memories onto younger generations as a crucial way of giving meaning to the past, and, therewith, in creating and sustaining identities (Hogervorst, 2010, Proske 2012, Macdonald, 2013; 200). Both practices respond to, and express, a perceived shift from war memory towards war history, although in different contexts of historical culture (Erll 2011; Assmann, 1999, 2006). It is through the digital revolution that, at the beginning of the 21st century, both of these practices have become intertwined; the shift from memory to history is accompanied by, or expressed by, a shift from collecting and preserving to disclosing digitized interviews for a wider audience, also for education (Bothe/Lücke, 2013; Barricelli, 2009, 2010).

My paper presents the first findings of a case study of my postdoc research project, that addresses the central question how online portals to digitized WWII eye witness testimonies are used in educational contexts – both formal education in history or civics, and heritage educational projects. The case study focuses on secondary school history teachers' conceptions of digitized video testimonies as educational resource. Therefore, a group was composed of teachers interested in WWII, which was asked to explore different online interview collections. Through participatory observation and individual interviews with participating teachers, data was collected on expectations, desires and experiences regarding the use of digitized video interviews in history classrooms. The analysis focused on two different themes: (1) Participants' conceptions of differences and similarities between live and digitized testimonies as educational resource; (2) Their experiences using specific online portals to interview collections. The latter will be the focus of my paper.

Participants were asked to explore two different online portals to different interview collections: Getuigenverhalen.nl ('eyewitness stories') and IWitness. Getuigenverhalen.nl is a Dutch portal hosted by the Netherlands Institute of War- Holocaust- and Genocidestudies, giving access to about 500 quite recently conducted video interviews (2007-2010), all in Dutch. The interviews address multiple WWII related topics such as resistance, daily life, persecution, and forced labor. The majority of these interviews is searchable at fragment-level as the transcripts and time-based key words that have been attributed are indexed and aligned with the video. The other portal is IWitness, the worldwide educational program of the USC Visual History Archive in Los Angeles. Through IWitness, in which a selected part (1,500 interviews) of the Shoah interview collection is made available online in an open, but supervised community of teachers and students. Twelve of these interviews are in Dutch. The video interviews can be watched and searched. Moreover, a video-editing tool enables users to select, annotate, and share video fragments, or to combine them with other fragments, photographs or information from the built-in encyclopedia. Because of IWitness' theoretical fundament in constructivism as a learning theory, and the fact that it actively invites teachers (and students) to create their own learning materials with/within the program, it is unclear whether this program can be easily implemented in non-American education systems and practices.

In Dutch history education, oral history is not a common practice; neither as a source of information that pupils learn to assess, nor as a practice in which pupils are trained. Regarding WWII however, there is a modest tradition of inviting eyewitnesses in classrooms, mostly in the weeks prior to the yearly commemorations of the liberation in May. Video interviews are hardly being used in Dutch history education at this point. There seems to be a transition period, in which institutions disclose their interview collections for still undefined audiences, which are mostly unaware of the existence of such collections and their online accessibility, and continue current practices. It is this transition that is the background of my postdoc research project. Other case studies focus on video interviews in WWII exhibitions and educational projects.

The aim of the postdoc research project is, first, to gain insight in contemporary historical culture, and specifically in the effects of the digital in transmitting and appropriating war memories across generations. For instance it would be important to know whether there is some kind of digital 'streamlining' of testimonies with specific features considered to be suitable for educational purposes. The same goes for the eye witnesses as culturally constructed figure, that is already very familiar to us through the numerous films, documentaries, news reports, and exhibitions (Kansteiner, 2015; Keilbach, 2012; Gries, 2012). Which characteristics of the narrators are perceived as necessary or relevant for letting students learn from their testimonies, and why exactly? And how about the perceived historical realism and authenticity of the testimonies – are we perhaps more critical when we do not encounter the narrators in person?

The second objective is to explore and think through the needs of the educational and heritage field as users of digitized oral history collections. This corresponds to research such as the CLARIAH project Oral History Today, that has made the numerous oral history collections in the Netherlands available for scholarly research, in close cooperation with both computer technologists and scholars in the humanities. (Scaglioa et al., forthcoming; Kemman et al., 2013; De Jong et al., 2011). It is relevant to expand the gained knowledge about the scholarly uses of oral history collections, examining how digital technology can be applied to turn interview collections into a network of knowledge relevant to multiple audiences, including teachers and students.

Bibliography

Andresen, K. et al. (eds.) (2015). *Es gilt das gesprochene Wort. Oral History und Zeitgeschichte heute*. Göttingen: Wallstein Verlag.

Apostolous, N. and Pagenstecher, C. (eds.) (2013). *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt*. Berlin: Metrolpol.

Assmann, A. and Brauer, J. (2011). Bilder, Gefühle, Erwartungen. Über die emotionale Dimension von Gedenkstätten und den Umgang von Jugendlichen mit dem Holocaust, *Geschichte und Gesellschaft* **37** (1): 72-103.

Barricelli, M. (2010). Kommemorativ oder kollaborativ? Historisches Lernen mithilfe digitaler Zeitzeugenarchive (am Beispiel des Visual History Archive). Alavi, B. (ed.), *Historisches Lernen im virtuellen Medium*. Heidelberg: Mattes Verlag, pp. 13-29.

Barricelli, M. (2009). Das Visual History Archive des Shoah Foundation Institute als geschichtskulturelle Objektivation und seine Verwendung im Geschichtsunterricht – ein Problemaufriss. H. Pandel, H. and Oswalt, V. (eds.), *Geschichtskultur. Die Anwesenheit von Vergangenheit in der Gegenwart*. Schwalbach im Taunus: Wochenschau Verlag, pp. 198-211.

Bothe, A. and Lücke, M. (2013). Shoah und historisches Lernen mit virtuellen Zeugnissen. Gautschi, P., Zülsdorf-Kersting M, Ziegler. B. (eds.), *Shoa und Schule. Lehren und Lernen im 21. Jahrhundert.* Zürich: Chronos Verlag, pp. 55-74.

Erll, A., (2011). *Memory in Culture*. Basingstoke and New York: Palgrave Macmillan.

Scagliola, S. and Jong, F. de (2014). Clio's talkative daughter goes digital. R. Bod, R. et al. (eds.) , *The Making of the Humanities, The Modern Humanities*, vol.**3** Amsterdam: Amsterdam University Press.

Hogervost, S. (2010). *Onwrikbare herinnering. Herinneringsculturen van Ravensbrück in Europa, 1945-2010*. Hilversum: Verloren.

Jong, F. de, Ordelman R. and Scagliola, S. (2011). Audio-visual Collections and the User Needs of Scholars in the Humanities, *Proceedings of Supporting Digital Humanities*, Copenhagen.

Kansteiner, W. (2014). Genocide memory, digital cultures, and the anesthetization of violence. *Memory Studies* **7** (4): 403–8.

Keilbach, J. (2012). Mikrofon, Videotape, Datenbank. Entwurf einer Mediengeschichte der Zeitzeugen. Sabrow, M. and Frei, N. (eds.), *Die Geburt des Zeitzeugen nach 1945*. Göttingen: Wallstein Verlag, pp. 281-99.

Kemman, M., et al. (2013) Talking with Scholars: developing a research environment for Oral History Collections. Demo paper. *Proceedings of the 2nd International Workshop on Supporting Users Exploration of Digital Libraries*, Malta.

Macdonald, S. (2013). *Memorylands. Heritage and identity in Europe today*. London and New York: Routlledge.

Stylometry on the Silver Screen: Authorial and Translatorial Signals in Film Dialogue

Agata Hołobut
aholobut@op.pl
Jagiellonian University, Krakow, Poland

Jan Rybicki
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland

Monika Woźniak
moniwozniak@gmail.com
Sapienza University of Rome, Italy

Stylometry based on quantitative analysis of linguistic features such as most frequent words, lemmata, or parts of speech, is a time- and research-proven tool in authorship attribution and plagiarism detection, and is now also used in more general literary studies as part of the distant reading revolution. It has been particularly successful in grouping long texts by their authors in both supervised and unsupervised machine learning tests – and the appeal of this material to stylometrists is understandable in that novels are easily available and easily definable chunks of linguistic (and literary) material, and, despite rumors on the death of the author, most readily associated by the general reader with a single creative figure. And when they are not, discovering the fingerprints of more than one hand in collaborative works is another favorite pastime of stylometrists.

While perhaps equally avidly studied, the authorship signal in drama is often more problematic. This is probably why the most famous question, that of Shakespearean authorship, is so complex and so hotly contested – as evident, for instance, in a fairly recent debate (Craig and Kinney, 2009; Vickers, 2011; Hoover, 2012). Other difficulties in this genre include the "codification of … literary discourse" in certain literary periods and the fact that the same authors might write drama both in prose and in verse (Schöch, 2013, 2014); also, it may be supposed that, as dramatists create their characters through dialogue, there is a more or less conscious effort on their part to differentiate their style. This last phenomenon has also been researched in novels (Burrows, 1987) and translations of novels and drama, and the results could be equally problematic (Rybicki, 2006, 2007, 2008).

Even more distortion may be expected in a somewhat similar genre, that of film and TV dialogue – and its textual reflection in intralingual subtitles and interlingual translations. The final shape of filmic speech ascribed to a given screenwriter can be influenced by other agents, such as directors or actors. It can be further transformed in the process of intralingual subtitling, especially that performed by "fansubbers", who do not necessarily reflect the exact dialogue spoken onscreen. This becomes even more of a problem in the case of interlingual translations, which by nature condense (subtitles, voice-over) or rework (dubbing) the original message, being at times anonymous versions of questionable quality ("fansubs").

Quantitative methods have already found their way into audiovisual translation research (Pérez-González 2014; Baños *et al.*, 2013), as exemplified by such projects as Pavia Corpus of Film Dialogue, used to examine sociolinguistic and pragmatic features of dubbed Italian (Freddi and Pavesi, 2009), or Forlixt1, a multimodal corpus which helps to investigate the interplay of verbal and non-verbal semiotics of the film (Valentini, 2006, 2008).

In comparison with the above attempts, our research was done on a specialized corpus of historical films and TV (mini)series. This choice was based on the assumption that the subgenre has unique characteristics which find reflection in film dialogue: namely, it authenticates the represented reality and it also adheres to the codes of realism existing in particular countries. This, in turn, made us look for the same phenomena in (English) originals and (Italian and Polish) translations, since translated dialogue, too, is shaped by stylistic necessities of the genre, culture-specific images of the past dominant in the target context, but also by norms and conventions of audiovisual translation in a given language/culture/country. The exact composition of the corpus is given in the table below:

Original	**Polish voice-over**	**Polish "official" subtitles**	**Polish fansubs**	**Italian dubbing**	**Italian "official" subtitles**	**Italian fansubs**
The Tudors Season 1 (2005)	+	+		+	+	+
The Tudors-Season 2 (2007)	+	+	+	+	+	+
Elizabeth I (2005) (miniseries)		+			+	
Elizabeth (1998)	+	+		+	+	+
Elizabeth. Golden Age (2007)	+	+		+	+	+
The Other Boleyn Girl (2008)		+		+	+	+
The Private Lives of Elizabeth and Essex (1939)			+	+	+	+
Anne of a Thousand Days		+		+		
Wolf Hall (2015)						+

From the point of view of film and audiovisual translation studies, our research project explores the concept of film/television genre and its distinctive features, focusing on the functions of film dialogue and linguistic/stylistic strategies used by screenwriters to fulfil them (Kozloff, 2000; Jaeckle, 2013). The first stage of our investigation consisted in a contrastive stylometric analysis of the extended Anglophone corpus, composed of both historical and non-historical film scripts, in order to verify our preliminary hypothesis about the genre-related specificity of film dialogue. We proceeded, then, to the analysis of parallel corpora of scripts in all available translations into Polish (voice-over, official and amateur subtitles) and Italian (dubbing, official and amateur subtitles). All this was done with several quantitative methods previously developed and used on other textual material, i.e. literary texts. In particular, frequencies of words from various frequency strata were compared between texts in each of the languages studied using the Delta procedure (Burrows, 2002). The analyses were performed with *stylo* (Eder et al., 2013), a package for R, the statistical programming environment (R Core Team, 2014), later also postprocessed with Gephi network analysis software (Bastian et al., 2009).

On the basis of these tests several observations could be made. As concerns the screenwriters, they tend to adapt the dialogues to the requirements of historical genre and the presented epoch. This is visible in Fig. 1, where the authorial signal seems to disappear whenever a given writer worked on two films and/or series set in different eras or belonging to a different, i.e. non-historical genre.

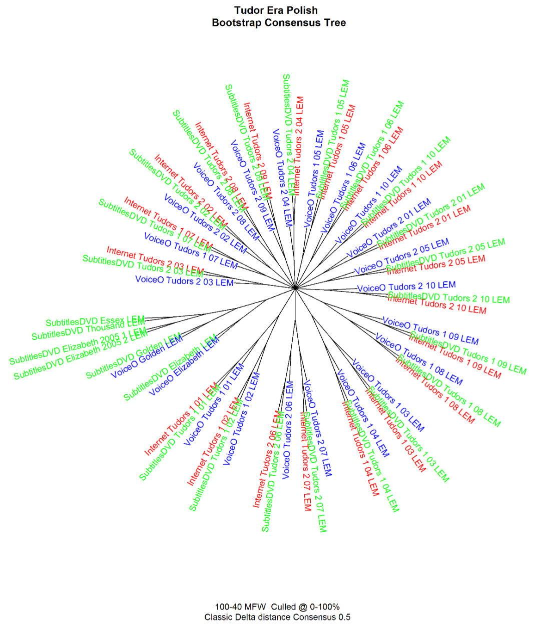


Figure 2. Bootstrap consensus tree for most frequent words in Polish subtitle and voice-over translations in a corpus of Elizabethan-era films and TV series.

As concerns audiovisual translations, we arrived at rather unexpected conclusions. We compared versions of individual episodes of TV series and films in Italian (dubbing, subtitles) and Polish (voice-over, subtitles) by analyzing frequencies of single words and part-of-speech 5-grams; the latter measure was a rough approximation

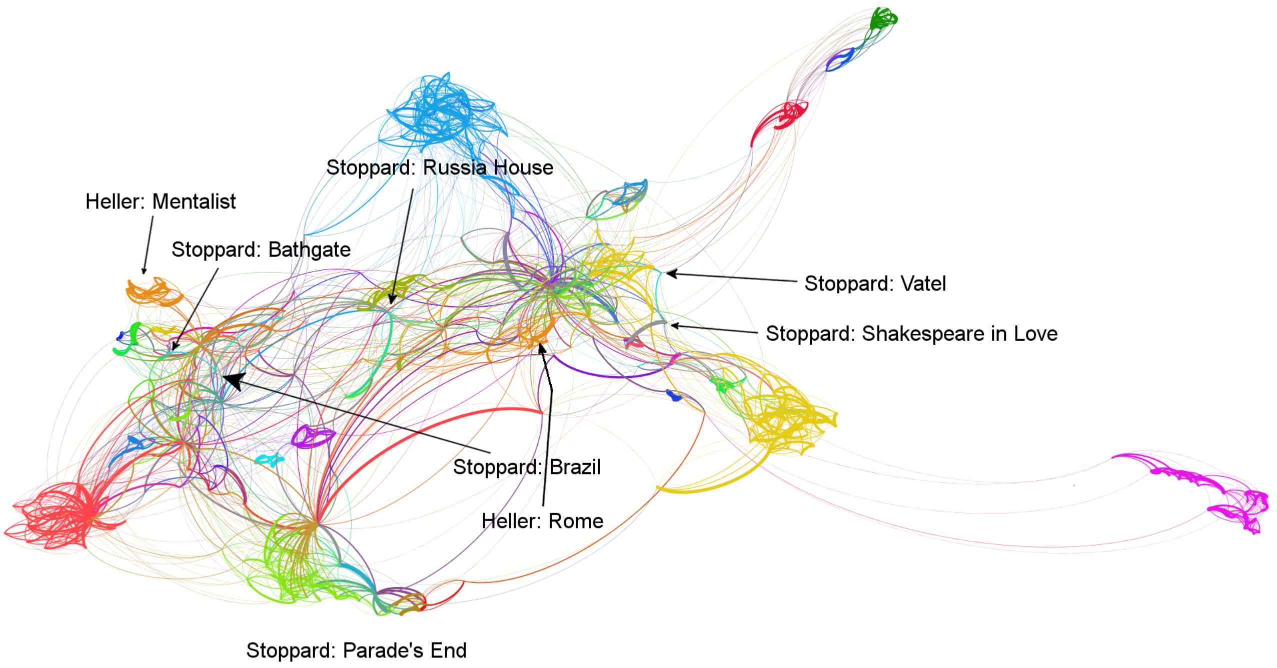


Figure 1. Network analysis diagram of historical and non-historical scripts.

of syntax (Górski et al., 2014). As far as Italy is concerned, we noticed an astounding uniformity of style regardless of technique, be it subtitles or dubbing: translations of individual episodes of TV series and films clustered together in analyses of both word and part-of-speech frequencies. By contrast, Polish subtitles and voice-over scripts of the same episodes clustered together for single-word frequencies, while the latter formed their own clusters in part-of-speech 5-gram analysis. This is shown in Fig. 2 and 3.

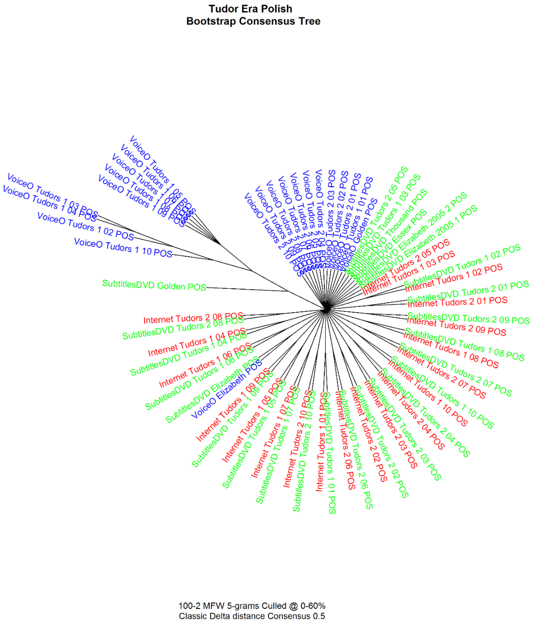


Figure 3. Bootstrap consensus tree for most frequent part-of-speech 5-grams in Polish subtitle and voice-over translations in a corpus of Elizabethan-era films and TV series.

Obviously, the similarities between dubbing and subtitles in Italian may stem from the fact that the latter are based on the former. However, the fact that even amateur subtitles, which usually are published before the release of the dubbed version, show a considerable affinity to dubbing, demonstrates high normalization of the formal language used in Italian historical films and television series.

All these results confirm our preliminary hypothesis that film genre influences the strategies of creating film dialogues and their translations. Although we believe that stylometric and computational analysis should not be the end in itself, it seems invaluable in audiovisual translation studies, encouraging closer qualitative analysis of the original and translated scripts. It invites further investigation of such issues as:

- the importance of cultural norms and conventions in film translation;
- significant intercultural differences in translation strategies used by subtitlers;
- complex relations between dubbing and subtitles, official and amateur subtitles, voice-over and subtitles
- cultural development of written / spoken language in a given country and the salient stylistic trends in audiovisual translation.

Acknowledgements

This research is part of project 2013/11/B/HS2/02890, supported by Poland's National Science Center.

Bibliography

Baños, R., Bruti, S. and Zanotti, S. (eds). (2013). *Perspectives: Studies in Translatology*. Special Issue: *Corpus Linguistics and Audiovisual Translation: In Search of an Integrated Approach*. **21**(4).

Bastian M., Heymann S. and Jacomy M. (2009). *Gephi: an Open Source Software for Exploring and Manipulating Networks*. International AAAI Conference on Weblogs and Social Media.

Burrows, J. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Oxford U. Press.

Burrows, J. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**: 267-87.

Craig, H., and Kinney, A. eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge U. Press.

Eder, M. (2015). Visualization in Stylometry: Cluster Analysis Using Networks. *Digital Scholarship in the Humanities*, **30**, first published online 3 December 2015, doi: 10.1093/llc/fqv061.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a Suite of Tools, in *Digital Humanities 2013: Conference abstracts*, University of Nebraska-Lincoln, pp. 487-89.

Górski, R., Eder, M. and Rybicki, J. (2014). Stylistic fingerprints, POS tags and inflected languages: a case study in Polish, in *Qualico 2014: Book of Abstracts*. Olomouc: Palacky University, pp. 51–53.

Freddi M, and Pavesi, M. (2009). The Pavia Corpus of Film Dialogue: Methodology and Research Rationale; in Freddi, M. and Pavesi M. (eds). *Analyzing Audiovisual Dialogue: Linguistic and Translational Insights*, Bologna: Clueb, pp. 95-100.

Hoover, D. (2012). The Rarer They Are, the More There Are, the Less They Matter. *Digital Humanities 2012: Conference abstracts*, University of Hamburg. Hamburg U. Press, pp. 218-221.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, **9**(6): e98679. doi:10.1371/journal.pone.0098679.

Jaeckle, J. (ed.). (2013). *Film Dialogue*. London & New York: Wallflower Press.

Jockers, M. (2013). *Macroanalysis. Digital Methods and Literary History*. Champaign: U. of Illinois Press.

Kozloff, S. (2000). *Overhearing Film Dialogue*. Berkley: University of California Press.

Pérez-González, L. (2014). *Audiovisual Translation Theories, Methods and Issues*. London and New York: Routledge.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien, http://www.R-project.org/.

Rybicki, J. (2006). Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's *Trilogy* and its Two English Translations. *Literary and Linguistic Computing* **21**(1), 91-103.

Rybicki, J. (2007). Twelve Hamlets: A Stylometric Analysis of Major Characters' Idiolects in Three English Versions and Nine Translations, in *Digital Humanities 2007: Conference Abstracts*, University of Illinois, Urbana-Champaign, p. 191.

Rybicki, J. (2008). Does Size Matter? A Re-examination of a Time-proven Method, in *Digital Humanities 2008: Conference abstracts*, University of Oulu, p. 184.

Schöch, C. (2013). Fine-Tuning our Stylometric Tools: Investigating Authorship and Genre in French Classical Drama, in *Digital Humanities Conference 2013*, Lincoln, Nebraska, USA.

Schöch, C. (2014). Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik, in Schöch, C. and Schneider, L. (eds). *Literaturwissenschaft im digitalen Medienwandel*, Mainz/Berlin: Philologie im Netz, pp. 130-57.

Valentini, C. (2006). A Multimedia Database for the Training of Audiovisual Translators. *JoSTrans: The Journal of Specialised Translation* 6. http://www.jostrans.org/issue06/art_valentini.php.

Valentini, C. (2008). Forlixt1: The Forli Corpus of Screen Translation: Exploring Macrostructures; in Chiaro, D. Heiss, Ch. And Bucaria, Ch. (eds). *Between Text and Image. Updating Research in Screen Translation*. Amsterdam & Philadelphia: John Benjamins, pp. 37-51.

Vickers, B. (2011). Shakespeare and Authorship Studies in the Twenty-First Century. *Shakespeare Quarterly* **62**: 106-42.

Identity Lenses in Analyzing Evolving Social Structures

John Robert Hott
jh2jf@virginia.edu
University of Virginia, United States of America

Worthy N. Martin
wnm@eservices.virginia.edu
University of Virginia, United States of America

Kathleen Flake
kathleen.flake@virginia.edu
University of Virginia, United States of America

In the effort to capture cultural dynamics, scholars have considered social networks, that is, a graph with people as nodes and their relationships as edges. These social networks are useful; however, to capture dynamics they must be considered over time. In the literature, Time-Varying Graphs (TVGs) have been defined (Aggarwal and Subbian, 2014; Casteigts et al., 2012; Casteigts et al., 2013). In our investigations, we have found benefit in defining TVGs with nodes as societal structures and people as the edges and then considering the dynamics of the societal structures evidenced in the TVGs (Hott et al., 2014; Hott et al., 2015). Here we consider two motivating applications for our extensions to TVGs: early Mormon marital structures and an arXiv.org citation network.

The societal structures represented in the marital and church structures of early Mormons in mid-1800s Nauvoo, Illinois, include binary, polygynous, and polyandrous marriages, as well as child and adult adoptions, and membership of individuals in the church organization hierarchy. In this time period the concept of "marriage" is in flux and part of our research is to consider various conceptualizations of "marriage" to better understand the relationship to the formation of the church structure. Each conceptualization we consider as a different "identity lens," a term we create to describe these different views.

We therefore define the *identity-lens function* that maps one evolving network to another evolving network. More specifically, given a TVG, $\mathcal{G} = (V, E, T, \rho, \zeta, \psi, \varphi)$ as defined in (Casteigts et al., 2012), the *identity-lens function* $f(i, \mathcal{G}) = \mathcal{G}_i$ maps the nodes and edges in $\mathcal{G}$ with a given identity definition to a new Time/Identity-Varying Graph (TIVG) $\mathcal{G}_i = (V_i, E_i, T, \rho_i, \zeta_i, \psi_i, \varphi_i)$. $\mathcal{G}_i$ is therefore the view of $\mathcal{G}$ under identity lens *i*.

In our marital network research (Hott et al., 2014; Hott et al., 2015), we represent marriages as the nodes, with the individuals connecting the marriages of their parents to their own marriages as adults. Every piece of this network is considered to be evolving, since marital relations change, new children are born, family members are adopted, and individuals change membership in the church organizational structure. Initially, this network $\mathcal{G} = \mathcal{G}_{binary}$ may be described as a binary-marriage network, in which each node depicts a marriage between two individuals and their biological children. This is one specific definition of node identity. However, we may examine this network in different levels of granularity: with different definitions of node identity. By extending the marriage definition to all individuals married to the same husband, we may redefine node identity to define polygynous marriages, creating $\mathcal{G}_{patriarchal}$. A related identity function that maps binary marriages to those with the same wife creates the TIVG $\mathcal{G}_{matriarchal}$.

Our second motivating example is the arXiv.org[1] citation network. ArXiv.org provides online open access to over one million cross-disciplinary papers, including papers in Physics, Mathematics, and Computer Science. We build a citation TVG from this dataset, linking authors as nodes based on the co-authorship of their papers. Similar to the Nauvoo application, we define multiple identity functions to map this TVG to multiple TIVGs. Under a node identity function combining authors within the same institution, we produce $\mathcal{G}_{institutional}$. Other node identity mappings include departmental affiliation $\mathcal{G}_{departmental}$, and $\mathcal{G}_{subject}$, in which authors are mapped to their subject areas.

Each of these TIVGs have characteristics that change over time. As we increase the complexity of the nodes through the use of identity lenses, we increase the dynamics of the characteristics, specifically those captured within the nodes. In the Nauvoo dataset, these characteristics include familial relationships among marriage members and church leadership positions held by the members of each marriage. Similarly, in the arXiv dataset, the characteristics include departmental and institutional collaboration. We want and need metrics that are sensitive to these changes within the evolving nodes as well as the overall evolving structure of the network. To capture and analyze these dynamics, we first define sampling methods to produce static graphs depicting the state of the TIVG during a fixed-size interval around each time point, then compute centrality measures over the graph across time for each identity lens. This process creates a distribution of the metric across time, which may then be compared between identity lenses. We conjecture that utilizing different-sized sampling intervals and comparing distributions across identity lenses will provide insights to understanding the TVG and the motivating application it describes.

We therefore define two methods to sample TIVG $\mathcal{G}_i$, in a λt-sized time interval around any given time-point *t* in *T*, to a static graph $G_i(t, \lambda t) = (V_i(t, \lambda t), E_i(t, \lambda t).)$They are given by the following node and edge set definitions:

1. The union of all nodes and edges extant at any time during the interval. ψ_i and ρ_i are the "presence" functions for nodes and edges, respectively, as defined in (Casteigts et al., 2012).

$$V_i(t, \lambda t) = \{v \in V_i \mid \exists t' \in [t - \lambda t/2, t + \lambda t/2] \Rightarrow \psi_i(v, t') = 1\}$$
$$E_i(t, \lambda t) = \{e \in E_i \mid \exists t' \in [t - \lambda t/2, t + \lambda t/2] \Rightarrow \rho_i(e, t') = 1\}$$

2. Only nodes and edges that exist throughout the entire interval.

$$V_i(t, \lambda t) = \{v \in V_i \mid \forall t' \in [t - \lambda t/2, t + \lambda t/2] \Rightarrow \psi_i(v, t') = 1\}$$
$$E_i(t, \lambda t) = \{e \in E_i \mid \forall t' \in [t - \lambda t/2, t + \lambda t/2] \Rightarrow \rho_i(e, t') = 1\}$$

As a simple example of these sampling methods, consider a correspondence network, where V is a set of individuals and E defines their correspondence; an edge connecting two individuals is present when a letter is in the mail between them. For an interval length, λt, of 1 year, the first sampling method would produce a graph containing connections between any individuals who corresponded by letter at any point that year. In comparison, the second sampling method would only leave connected those individuals with constant communication for the entire year.

Using the sampling methods above, we measure characteristics at time points throughout the lifetime of $\mathcal{G}_i$ and thereby evidence the dynamics. The harmonic centrality, C_H, is an indication of how close the nodes are to each other, while the betweenness centrality, C_B, is indicative of how interconnected the nodes are within the graph. They are defined as

$$C_H(G_i(t, \lambda t)) = \frac{\sum_{v_j \in V_i(t,\lambda t)} \left(\max_{\forall u \in V_i(t,\lambda t)} (C^*_H(u, G_i(t, \lambda t))) - C^*_H(v_j, G_i(t, \lambda t)) \right)}{[(|V_i(t, \lambda t)| - 1)(|V_i(t, \lambda t)| - 1)]/(2|V_i(t, \lambda t)| - 3)} \text{ and}$$

$$C_B(G_i(t, \lambda t)) = \frac{2 \sum_{v_j \in V_i(t,\lambda t)} \left(\max_{\forall u \in V_i(t,\lambda t)} (C^*_B(u, G_i(t, \lambda t))) - C^*_B(v_j, G_i(t, \lambda t)) \right)}{(|V_i(t, \lambda t)| - 1)^2(|V_i(t, \lambda t)| - 2)},$$

where $C^*_H(v_j, G_i(t, \lambda t))$ and $C^*_B(v_j, G_i(t, \lambda t))$ are the harmonic and betweenness point-centrality measures (Wasserman and Faust, 1994) for a given node $v_j \in V_i(t, \lambda t)$, respectively. For brevity, we will define here only C^*_H, using Boldi and Vigna's harmonic centrality definition (Boldi and Vigna, 2014), as

$$C^*_H(x, G_i(t, \lambda t)) = \sum_{\substack{d(y,x)<\infty \\ y \neq x \in V_i(t,\lambda t)}} \frac{1}{d(y, x)},$$

where $d(y, x)$ is the distance between $y, x \in G_i(t, \lambda t)$. As a concrete example of this measure, consider the graphs in Figures 1 and 2. In the graph of Figure 1, node A acts as the central connection point, or hub. The hub has the shortest distance to any node and therefore high harmonic point-centrality, C^*_H = 6. Other nodes must traverse at most two edges to reach any other node, giving them C^*_H = 3.5. The overall measure for this graph is C_H = 4.58. In comparison, the graph in Figure 2 has nodes that are distant from most of the other nodes, e.g., node A has C^*_H = 2.45 leading to harmonic centrality, C_H = 1.02. These two graphs demonstrate that the harmonic centrality of the graph is inversely related to the overall "closeness" of the nodes.

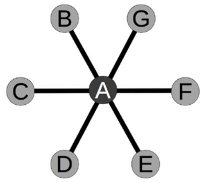


Figure 1. A star graph, with nodes shaded based on relative point-centrality, which has harmonic centrality C_H = 4.58

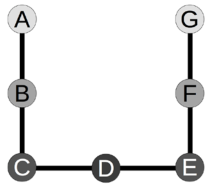


Figure 2. A linear graph, with nodes shaded based on relative point-centrality, which has harmonic centrality C_H = 1.02

Allowing *t* to range over the entire lifespan of $\mathcal{G}_i$ and considering multiple sizes for our λt interval, we generate distributions of the metric across time and with differing levels of temporal granularity. These distributions give a picture of the dynamics occurring within each TIVG. By comparing the metrics across TIVGs under different identity functions for the same TVG, we hope to more fully capture the dynamics of and understand the original evolving network, and provide insights into the motivating application at hand.

We have therefore defined a new conceptualization of Time-Varying Graphs, specifically the identity-lens function and resulting Time/Identity-Varying Graphs under each identity mapping. We also defined methods for sampling the TIVGs into series of measurable static graphs and provided metrics over those representations. At the conference, we will present visualizations that represent each of our applications from the various perspectives, as well as the findings of these measures: to better understand the definition of marriage in Nauvoo and its relation to church formation, and to illuminate patterns in author and departmental co-citations.

Bibliography

Aggarwal, C. and Subbian, K. (2014). Evolutionary network analysis: A survey, *ACM Computing Surveys (CSUR)*, **47**(1): 10.

Boldi, P. and Vigna, S. (2014). Axioms for centrality, *Internet Mathematics* **10**(3-4): 222–62.

Casteigts, A., et al. (2013). *Expressivity of time-varying graphs, Fundamentals of Computation Theory*, Springer, pp. 95–106.

Casteigts, A., et al. (2012). Time-varying graphs and dynamic networks, International Journal of Parallel, *Emergent and Distributed Systems* **27**(5): 387–408.

Hott, J. R., Martin, W. N. and Flake, K. (2014). *Evolving social structures: Networks with people as the edges*, Digital Humanities Forum, University of Kansas.

Hott, J. R., Martin, W. N. and Flake, K. (2015). Visualizing and analyzing identity classes in evolving social structures, *Chicago Colloquium on Digital Humanities and Computer Science*.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Vol. 8, Cambridge University Press.

Notes

[1] *http://www.arxiv.org*

Exploring the Rules of Rhyme: Operationalizing Historical Poetics

Natalie M. Houston
Natalie_Houston@uml.edu
University of Massachusetts-Lowell, United States of America

Rhyme is a key feature of many verse forms used throughout the history of English poetry. In its simplest forms, rhyme in English poetry can be understood as a connection between words at the end of lines of poetry, due to the similarity in the sounds of their final syllables. Rhyme suggests and enacts relationships between sound and sense in poetry: "the equivalence of the rhyme syllables or words on the phonic level implies a relation or likeness or difference on the semantic level" (Brogan et al., 2012). Internal rhyme can also connect words located elsewhere in the poetic line, although that is less common in English poetry. Rhyme connects lines of poetry through sound patterns which contribute to the tone, pace, and emotional effects of a poem. Because a common printing convention in the nineteenth century indented lines of poetry to match the rhyme scheme, rhyme was also a visual feature of printed poems, perceptible to the reader's eye and cognition. Rhyme, along with meter, contributes to the memorable qualities of verse, and to the popularity of humorous forms like the limerick. Rhymed verse forms, such as the sonnet and ballad, were very popular in nineteenth-century British poetry, alongside unrhymed blank verse. Rhyme was used for different ideological and aesthetic purposes, as in the working-class ballads of the 1840s or Decadent revivals of French forms like the triolet and villanelle.

But the rules of rhyme are not fixed. Throughout the nineteenth century, poets and critics debated which kinds of rhymes were allowed in the best poetry. In particular, some critics argued that imperfect rhyme (also called near rhyme) was allowable because so many poets used such rhymes (such as "love" and "prove"), while others argued for a more restrictive definition of perfect rhyme. These debates about rhyme are part of larger nineteenth-century discourses about prosody, the patterning of sound in poetry. Recent work in nineteenth-century studies has focused on "historical poetics," the examination of nineteenth-century theories of prosody and their cultural impact (Hall 2011, Martin 2012, Rudy 2009). Such scholarship complicates transhistorical definitions of lyric poetry and its formal features by demonstrating how "prosody provided a way of thinking, a method of protest, of scientific argument and investigation, of negotiating gender, class, and national structures" (Martin and Levin, 2011: 153). But too often this work in historical poetics remains focused on theory separate from poetic practice: Yopie Prins, for instance, says bluntly that "practical application is not the point of historical poetics. There are other, more interesting questions" which for her encompass the relationship of prosody to larger philosophical and scientific discourses (Prins, 2008: 233). This paper offers an alternative, computational approach to historical poetics that will not only further our understanding of nineteenth-century theories of rhyme but also of their relationship to actual poetic practice.

Rhyming dictionaries, which serve as a resource to writers seeking rhymes for their compositions, serve as an importance source for understanding the changing rules of rhyme in the nineteenth century. In the eighteenth century, Edward Bysshe's *The Art of English Poetry* (1702) offered a small dictionary of rhymes, supplanted later in the century by John Walker's *A Dictionary of the English Language,* *Answering at once the Purposes of Rhyming, Spelling, and Pronouncing*, first published in 1775 with 41,000 words included in the rhyme dictionary. Walker's dictionary was expanded and reprinted many times throughout the nineteenth and twentieth centuries, and it is mentioned in the letters and papers of many writers and poets as a standard text. But in the second half of the nineteenth century the popularity of poetry writing as a leisure activity among the growing middle class led to the publication of many new rhyme dictionaries, each of which offered different definitions of acceptable rhymes. These included J. E. Carpenter's *A Handbook of Poetry* (1868); Tom Hood's *The Rules of Rhyme* (1869), later republished as *The Rhymester; Or, the Rules of Rhyme*; the American writer Samuel W. Barnum's *A Vocabulary of English Rhymes, Arranged on a New Plan* (1876); John Longmuir's *Rhythmical Index to the English Language* (1877); R. F. Brewer's *Orthometry: A Treatise on the Art of Versification* (1893); and Andrew Loring's *The Rhymer's Lexicon* (1905).

This paper presents my current work in progress in operationalizing these nineteenth-century rhyme dictionaries as R scripts that identify rhymes in poetic texts based on the rules of a specific dictionary. The entries in these rhyme dictionaries consist of rhyme syllables that serve as the entry headings, followed by a list of homophone syllables, and then examples of words that rhyme with the entry heading. The rhyme syllables and words listed may be subdivided into categories, such as perfect rhymes and imperfect or "allowable" rhymes. Many of the dictionaries provide supporting evidence for the use of imperfect rhymes in the form of quotations from the works of British poets. For this project, all of the rhyme syllables, rhyme words, and rhyme categories are stored in csv files for each dictionary, organized by entry headings. In the interest of fidelity to the original historical dictionaries, and to facilitate comparative analysis, conflicting entries have not been normalized. Thus one dictionary might give very different rhyme syllables or words for a given entry than those listed in another; it is precisely this kind of historical variation in the rules of rhyme that this project allows us to explore. The R package data.table facilitates querying and joining data files for multiple dictionaries so as to compare their entries for the same rhyme syllable.

The technical section of the paper describes the key components of these scripts:

- the last word of each line of the poem is extracted from the text and converted to reverse spelling;
- a series of regular expressions are used to enact the specific instructions each rhyme dictionary provides for its users to identify the rhyme syllable to look up, such as locating the vowel that precedes the consonant(s) of the final syllable ("ame"), unless the consonant(s) are preceded by a dipthong, in which case the first of the two dipthong vowels begins the rhyme syllable ("ound");
- a hash lookup for each rhyme syllable to the selected dictionary returns the perfect rhyme syllables, perfect rhyme words, imperfect rhyme syllables (if given), and imperfect rhyme words (if given);
- matches among these returned values and the other rhyme syllables and words in the poem are sought; when located, a capital letter is used to mark the rhyme pattern, as is traditional in literary criticism (ABAB, etc);
- for those dictionaries that include imperfect rhymes, the ratio of perfect to imperfect rhymes in the poem is also recorded.

This project expands our understanding of nineteenth-century rhyme theory and practice in three ways: by enabling the identification of rhyme patterns in large sets of texts, we can expand our understanding of historical verse writing practices beyond canonical literary texts; by operationalizing the rules instantiated in different dictionaries, we can compare how they would have evaluated the rhymes chosen by particular poets; and by creating a database of the words included in these dictionaries, we can examine the consistency and variation of the recommended rhymes.

The final section of the paper presents a case study in the application of these scripts to literary texts, an analysis of the rhyme patterns used in the 1274 poems included in Edmund Clarence Stedman's *A Victorian Anthology 1837-1895* (1895), as a sample dataset of poetry produced during the time when these rhyme dictionaries were circulating and when these rhyme theories were debated. Examining the frequency of different rhyme patterns and the vocabulary of rhyme common in Victorian poetry opens up new paths for analyzing rhyme practice, rhyme theory, and the lexical fields generated by the use of rhyme. This project thus contributes to current work in nineteenth-century studies and historical poetics by computationally analyzing the theory offered by rhyme dictionaries with the actual practices of nineteenth-century poetry.

Bibliography

Brogan, T.V.F., et al. (2012). Rhyme. In Cushman, S. and Cavanagh, C. (eds), *Princeton Encyclopedia of Poetry and Poetics*. Princeton: Princeton University Press, pp. 1182-92.

Hall, J. (ed). (2011). *Meter Matters: Verse Cultures of the Long Nineteenth Century*. Athens: Ohio University Press.

Martin, M. (2012). *The Rise and Fall of Meter: Poetry and English National Culture, 1860-1930*. Princeton: Princeton University Press.

Martin, M. and Levin, Y. (2011). Victorian Prosody: Measuring the Field. *Victorian Poetry*, **49**(2): 149-60.

Prins, Y. (2008). Historical Poetics, Dysprosody, and 'The Science of English Verse'. *PMLA* **123**(1): 229-34.

Rudy, J. (2009). *Electric Meters: Victorian Physiological Poetics*. Athens: Ohio University Press.

From Keyword Search To Discourse Mining - The Meaning Of Scientific Management In Dutch Vocabulary, 1900-1940

Pim Huijnen
p.huijnen@uu.nl
Utrecht University, Netherlands, The

Juliette Lonij
juliette.lonij@kb.nl
Koninklijke Bibliotheek, The Hague, Netherlands, The

In this paper we present a technique to enable the historical study of ideas instead of words. It aimed at assisting humanities scholars in overcoming the limitations of traditional keyword searching by making use of context-specific dictionaries. The elaboration of this technique was the result of a successful collaboration between the History Department of Utrecht University (UU) and the Research Department of the Koninklijke Bibliotheek, National Library of the Netherlands (KB), executed by the authors of this paper during Huijnen's period as Researcher-in-residence at the KB in 2015.

The aim of this collaborative project was twofold: first, to create a method for dictionary extraction from a representative text corpus, based on existing methods and algorithms. Second, to find a way of executing dictionary searches in the KB's digitized newspaper archive and visualizing the results. Both components of the project were tested and evaluated by means of a case study on the impact of American scientific management theories in the Dutch public sphere during the first half of the 20th Century. Using the approach described here, we were able to discover and analyze shifts in the way the modernization of Dutch business and economy was discussed during this period. We would not have been able to achieve the same results by means of traditional historical scholarship alone.

Historical newspapers have traditionally been popular sources to study public mentalities and collective cultures within historical scholarship. At the same time, they have been known as notoriously time-consuming and complex to analyze. The recent digitization of newspapers and the use of computers to gain access to the growing mass of digital corpora of historical news media are altering the historian's heuristic process in fundamental ways.

The large digitization project the Dutch National Library currently runs can illustrate this. Until now, the KB has made publicly available over 80 million historical newspaper articles from the last four centuries. Researchers (as well as the wider public) are able to do full-text searches in the entire repository of articles through the KB's own online search interface Delpher (http://www.delpher.nl/kranten). Instead of manually skimming through a selected numbers of editions or volumes this functionality allows for the searching of particular (strings of) keywords within the entire corpus. As basic as it may seem, full-text searching completely overturns the way in which historians are used to approach newspapers. Instead of the successive top-down selections historians traditionally made in order to gradually isolate potentially interesting material, keyword searching treats the corpus as a singular bag of words and, therefore, enables researchers to immediately dive into the texts that meet their search criteria (Nicholson, 2013).

At the same time, keyword searching has some serious shortcomings for the use in (cultural) historical research. Historians commonly work with texts, but are rarely interested in language per se. Rather, they use written or spoken sources (be it correspondence, literature, diaries, or news media) to gain access to past cultures, ideas, or mentalities. The things that historians are mostly interested in, are often not made explicit (e.g. the Enlightenment attitude, generational conflicts) and difficult to abstract into singular keywords (modernity, secularization). Doing historical research with keyword searching is like painting a canvas using felt-tip pens: it loses every inch of subtlety.

We have successfully developed a technique of dictionary extraction and searching to address this problem. The use of dictionaries is able to bring greater subtlety and diversity into digital historical scholarship. The more elaborate these dictionaries are, the more they overcome the contingency that comes with the use of singular keywords in search strategies. Several research projects that have incorporated the use of highly domain- and time-specific word-lists ('dictionaries'), have already shown this. Text classification algorithms, for example, have helped find the most obvious indicator words for articles about strikes in the Dutch newspaper corpus (Van den Hoven et al., 2010). Implicit dictionaries based upon the MALLET (http://mallet.cs.umass.edu) package's topic modeling functionality has assisted in finding Darwinian motives in Danish literature (Tangherlini and Leonard, 2013). Topic modeling was also used in building a neoliberalism dictionary to study Colin Crouch's post-democracy thesis in German historical newspapers (Wiedemann et al., 2013; Wiedemann and Niekler, 2014).

From the wide variety of techniques scholars have developed to build and use dictionaries, this project found most inspiration in the topic modeling-based method of the ePol Projekt (Wiedemann and Niekler, 2014). However, rather than aiming at building an optimal infrastructure for dictionary extraction of our own, based on existing techniques, our project centered around practical usability. We sought to develop a (set of) tool(s) for working with dictionaries tailored to the computational expertise to be expected from, but also the specific needs of professional historians (and humanities scholars in general). One of the aims of the KB's Researcher-in-residence program, in addition, is that resulting tools and techniques are usable by the wider public searching the National Library's databases of historical newspapers, periodicals, and books. Our code is fully open source and can be found on GitHub (https://github.com/jlonij/keyword_generator). The ways in which we have tried to meet the specific demands this posed, can, in our view, be seen as exemplary for any Digital Humanities project aimed not at building highly specialized tools for individual projects, but at combining scholarly standards with the goal of generic usability.

There are a number of ways we have accounted for the targeted user groups in the development of our dictionary extraction and search techniques. On the one hand, for example, we aimed at agility and flexibility at the expense of the deployment of exhaustive computational means. Our algorithm is able to extract a dictionary of flexible length from a given source input of text files within minutes. Because the technique is intended for exploratory use, it is essential that iterations and experimentations are stimulated. Requiring too many preprocessing steps or demanding too much time would be counterproductive.

On the other hand, meeting the demands of tool criticism was crucial in every step of this project. Therefore, the risk of blackboxing was avoided wherever we could, while at the same time granting the user-expert as much control as possible. By varying the command, users decide over the segmentation of the source corpus, the number of topics to be generated, the number of words to be contained per topic, as well as the number of dictionary words required. Moreover, users may flexibly choose between Gensim's (https://radimrehurek.com/gensim) and MALLET's implementation of LDA, as well as a straightforward tf-idf implementation. When making use of one of the topic modeling packages, users are, just before the process of dictionary generation, given the option of excluding any number of (irrelevant) topics of choice from the equation.

An evaluation in terms of generic precision and recall for any of the variables is, in our view, contradictory to the principle of flexibility. Instead, we evaluated and improved the dictionary extraction by comparing automatically generated dictionaries with ones that were built manually, based on domain knowledge. Comparing the results of searches with different dictionaries in the KB's digitized newspaper archive was used as an additional evaluation method: dictionaries could be compared in terms of the ranking of some key articles about a particular topic, since the archive's Solr (http://lucene.apache.org/solr) search engine scores the results of an OR-query (the search string, in which we expressed the dictionaries) on the basis of the number of times query words appear in an article, amongst other things. The case study that was used to test, evaluate, and apply the tools and techniques under development was the impact of American scientific management ideas in the Dutch public media before WWII.

On the basis of this case study we would in our presentation like to show how our implementation of dictionary extraction, search, and visualization can assist the scholarly historical study of digital corpora in general. By visualizing the search results from different dictionaries we are able to show shifting discourses in historical news media. Plotting the number of articles containing a user-specified number of words from any given dictionary, we can present trends in discourse-specific vocabulary usage over time. Whereas existing historiography, for example, suggests a continuing use of scientific management vocabulary in the Netherlands since its introduction in the 1910s, our project presents a more differentiated picture. Dictionary searches in the KB's newspaper corpus show how the use of words in public media connected to the sphere of scientific management (based on context- specific literature) waned after the WWII and how they made room for a new vocabulary belonging to a new era.

At the same time, this case study illustrates how digital techniques like ours bring about conceptual innovations in the study of history. After all, our case study shows that (combinations of) ordinary words (in this instance, for example, 'time', 'work', or 'supervision') are more distinguishing to trace discursive discontinuities than the 'big' words (like 'taylorism' or 'neoliberalism') that historians traditionally have focused on.

Bibliography

Nicholson, B. (2013). The Digital Turn, *Media History*, **19**: 59-73.

Tangherlini, T. R. and Leonard, P. (2013). Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research, *Poetics*, **41**: 725-49.

Van den Hoven, M., Van den Bosch, A. and Zervanou, K. (2010). Beyond Reported History: Strikes That Never Happened. *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts,* Vienna, pp. 20-28.

Wiedemann, G., Lemke, M. and Niekler A. (2013). Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentation in der Bundesrepublik Deutschland 1949-2011, *Zeitschrift für Politische Theorie*, **4**: 99-115.

Wiedemann, G. and Niekler, A. (2014). Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries. *Terminology and Knowledge Engineering,* Berlin, June 2014. https://hal.archives-ouvertes.fr/hal-01005879 (accessed 2 March 2016).

Publishing Second World War History as Linked Data Events on the Semantic Web

Eero Hyvönen
eero.hyvonen@aalto.fi
Aalto University, Finland

Erkki Heino
erkki.heino@aalto.fi
Aalto University, Finland

Petri Leskinen
petri.leskinen@aalto.fi
Aalto University, Finland

Esko Ikkala
esko.ikkala@aalto.fi
Aalto University, Finland

Mikko Koho
mikko.koho@aalto.fi
Aalto University, Finland

Minna Tamper
minna.tamper@aalto.fi
Aalto University, Finland

Jouni Tuominen
jouni.tuominen@alto.fi
Aalto University, Finland

Eetu Mäkelä
eetu.makela@aalto.fi
Aalto University, Finland

Second World War on the Semantic Web

Data about wars is typically heterogeneous, distributed in the data silos of the fighting parties, multilingual, and often controversial depending on the political point of view. It is therefore hard for the historians to get a global picture of what has actually happened, to whom, where, when, and how. We argue that Semantic Web and Linked Data technologies are a very promising approach for modeling, harmonizing, and aggregating data about war history. Our goal is to make it possible, for both historians and laymen, to study history in a contextualized way where linked datasets enrich each other. The paper presents the in-use WarSampo[1] system, where massive collections of heterogeneous data about the (Finnish) history of the Second World War are harmonized using an event-based approach, and provided as a Linked Open Data service for applications to use. As a use case, a semantic portal WarSampo providing six different perspectives to the war based on events is presented.

There are several projects publishing data about the World War I on the web, such as Europeana Collections 1914–1918[2], 1914–1918 Online[3], WW1 Discovery[4], Out of the Trenches[5], CENDARI[6], Muninn[7], and WW1LOD (Mäkelä et al., 2015). War history makes a promising use case for Linked Data because war data is heterogeneous, distributed in different countries and organizations, and written in different languages (Hyvönen, 2012). Using metadata, also different opinions and conflicting information about the war can be represented.

Many websites also publish information about the World War II, the largest global tragedy in human history, such as the World War II Database[8] to name just one. However, such portal data is typically meant for human consumption, and there are only few works that deal with machine readable data about the WW2 for applications to use (Collins et al., 2005; de Boer et al., 2013).

Our work contributes to the related research above by initiating and fostering large scale LOD publication of WW2 data on the web, based on event-based data modeling. The idea is to publish Linked Open data, aggregated from distributed data silos, for Digital Humanities applications to use. In our case study, the data is related to the Finnish Winter War 1939–1940 against the Soviet attack, the Continuation War 1941–1944, where the occupied areas of the Winter War were temporarily regained by Finns, and the Lapland War 1944–1945, where the Finns pushed the German troops away from Lapland.

We first present and discuss the data modeling approach developed for the WarSampo LOD service. After this an application of the data, the WarSampo semantic portal, is presented where events are linked to related resources, such as photos, persons, and historical places. In conclusion, lessons learned are discussed and directions for further research pointed out.

The data service: modelling war events as Linked Data

Dataset	Name	Providing organization	Size
1	Casualties of WW2	National Archives	95,000 death records
2	War diaries	National Archives	26,500 war diaries of army units
3	Photos & films	Defence Forces & Military Museum	160,000 photos & films
4	Kansa taisteli magazine articles	Bonnier & The Assoc. for Military History in Finland	3,360 articles of veteran soldiers
5	Karelian map names	National Land Survey	35,000 places of the annexed Karelia
6	Karelian maps	National Land Survey	47 war time maps of Karelia aligned on modern maps
7	Organization cards	National Archives	War events of ca. 500 army units
8	War time events	Books etc.	11 300 war events in 1939-45

Table 1. Central datasets published and linked in WarSampo.

Data

The project deals initially with the datasets presented in Table 1. The casualties data (1) includes data about the deaths in action during the wars. War diaries (2) are digitized authentic documentations of the army unit actions in the frontiers. Photos and films (3) were taken during the war by the troops of the Defense Forces. The Kansa taisteli magazine (4) was published in 1957–1986; its articles contain mostly memories of the men that fought on the fronts. Karelian places (5) and maps (6) cover the war zone area in pre-war Finland that was finally annexed by the Soviet Union. Organization cards (7), written after the war, document events of military units during the war. War time events (8) extracted from various publications include, e.g., battles and political incidents. The data, over 5 million triples in total, has been transformed into RDF from database dumps, spreadsheets (CSV), and by applying OCR to documents. Named Entity Recognition (NER) techniques were used to link texts to data, e.g., to identify and disambiguate persons and places in the magazine articles and captions of the photos. In addition, new datasets are planned to be included in the system, such as the Finnish Broadcasting Company YLE's audio and film material recorded during the war or related to it ("Living Archive"), and a database of prisoners of war.

Metadata models

CIDOC CRM[9] is used as the harmonizing basis for modeling data, with events providing the semantic glue for data linking (Doerr, 2003). Our data model for WW1, presented in (Mäkelä et al., 2015), is used as the metadata model to start with. The model and data is documented at the data service[10].

Domain ontologies

The data is annotated using a set of domain ontologies, including: 1) an ontology of the troops and their hierarchies, 2) persons with their ranks and roles, 3) place ontology of historical places, 4) event ontology of battles, politics, and other war time incidents, 5) an ontology of time periods, and 6) a subject matter ontology. Ontologies on objects such as weapons, aircraft, and vessels remain topics of possible future work.

All WarSampo datasets are available as Linked Open Data (LOD) at the "7-star" Linked Data Finland service[11] (Hyvönen et al., 2014), based on Fuseki[12] establishing the SPARQL endpoint, and with a Varnish Cache[13] frontend for dereferencing URIs.

Application: perspectives to war history

The idea of the WarSampo portal is to provide a variety of different perspectives to war history. There are six perspectives available: Events, Persons, Army Units, Places, Kansa taisteli Magazine Articles, and Casualties (Hyvönen et al., 2016). The idea is that the perspectives enrich each other based on data linking.

Figure 1 illustrates the WarSampo Events perspective application to the WarSampo data. Events are displayed on a map, (a) in Fig. 1, and on a timeline (b) that shows here some events of the Winter War. When the user clicks on an event, it is highlighted (c), and the historical place, time span, type, and description for the selected event are displayed (d). Photographs related to the event (e) are also shown. The photographs are linked to events based on location and time. Furthermore, information about casualties during the time span visible on the timeline is shown alongside the event description (f), and the map (a) features a heatmap layer for a visualization of these deaths.

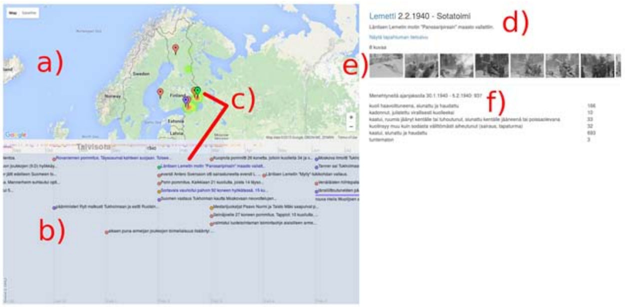


Figure 1. Event perspective featuring a timeline and a map.

The events can also be found and visualized through other perspectives. For example, in the Army Units perspective, the events in which a unit participated can be viewed on maps and in time, providing a kind of graphical activity summary of the unit. In the casualties perspective, military units of the dead soldiers are known, making it possible to sort out and visualize the personal war history of the casualties on historical maps that come from a yet another dataset in WarSampo.

The WarSampo semantic portal[14] was published Nov 27, 2015 at and has had tens of thousands of users. It is implemented solely on the SPARQL endpoint of the WarSampo LOD data service.

Discussion

Our first experiments, as illustrated in Section 3, suggest that heterogeneous datasets of war history really can be interlinked with each other through events in ways that provide useful insights for the historians. We have also learned that even in the rural northern parts of Europe, massive amounts of WW2 data can be found. We have initially dealt with tens of thousands of people killed in action. However, there is also data available about hundreds of thousands of soldiers who survived the war. In addition to historians, WarSampo data is very interesting to the laymen, too: every soldier's history is of interest at least to, e.g., his/her relatives. Managing the data, and providing it for different user groups, suggests serious challenges when dealing with, e.g., the war in the central parts of Europe, where the amount of data is orders of magnitude larger than in Finland, multilingual, and distributed in different countries. For example, solving entity resolution problems regarding historical place names and person names can be hard. However, it seems that Linked Data is a promising way to tackle these challenges.

Our work[15] is funded by the Ministry of Education and Culture and Finnish Cultural Foundation. Wikidata Finland project financed the alignment of the historical maps in WarSampo.

Bibliography

De Boer, V., van Doornik, J., Buitinck, L., Marx, M. and Veken, T. (2013). Linking the kingdom: enriched access to a historiographical text. In: *Proc. of the 7th International Conference on Knowledge Capture* (KCAP 2013). Association of Computing Machinery, New York, pp. 17–24.

Collins, T., Mulholland, P. and Zdrahal, Z. (2005). Semantic browsing of digital collections. In: *Proc. of the 4th International Semantic Web Conference* (ISWC 2005). Springer–Verlag.

Doerr, M. (2003). The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine, **24**(3): 75–92.

Hyvönen, E. (2012). *Publishing and using cultural heritage linked data on the semantic web.* Morgan & Claypool, Palo Alto, CA, USA.

Hyvönen, E., Lindquist, T., Törnroos, J. and Mäkelä, E. (2012). History on the semantic web as linked data – an event gazetteer and timeline for World War I. In: *Proc. of CIDOC 2012 – Enriching Cultural Heritage.* http://seco.cs.aalto.fi/publications/2012/hyvonen-et-al-ww1-cidoc-2012.pdf

Hyvönen, E., Tuominen, J., Alonen, M. and Mäkelä, E. (2014). Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: *The Semantic Web: ESWC 2014 Satellite Events*, Revised Selected Papers. Springer–Verlag, pp. 226–30.

Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J. and Mäkelä, E. (2016). WarSampo data service and semantic portal for publishing linked open data about the second world war history. In: *Proceedings of the 13th Extended Semantic Web Conference* (ESWC 2016). Springer–Verlag, forth-coming.

Mäkelä, E., Törnroos, J., Lindquist, T. and Hyvönen, E. (2015). World War I as Linked Open Data (2015). http://www.seco.tkk.fi/publications/submitted/makela-et-al-ww1lod.pdf, submitted for review.

Notes

[1] “Sampo" is a magical artifact in Finnish mythology that brought good fortune to its owner.

[2] http://www.europeana-collections-1914-1918.eu

[3] http://www.1914-1918-online.net

[4] http://ww1.discovery.ac.uk

[5] http://www.canadiana.ca/en/pcdhn-lod/

[6] http://www.cendari.eu/research/first-world-war-studies/

[7] http://blog.muninn-project.org

[8] http://ww2db.com

[9] http://cidoc-crm.org

[10] http://www.ldf.fi/dataset/warsa/

[11] http://www.ldf.fi

[12] http://jena.apache.org/documentation/serving\data/

[13] https://www.varnish-cache.org

[14] The application is in use at http://sotasampo.fi.

[15] See the project homepage http://seco.cs.aalto.fi/projects/sotasampo/en/.

Contextualizing Historical Places in a Gazetteer by Using Historical Maps and Linked Data

Esko Ikkala
esko.ikkala@aalto.fi
Aalto University, Finland

Jouni Tuominen
jouni.tuominen@aalto.fi
Aalto University, Finland

Eero Hyvönen
eero.hyvonen@aalto.fi
Aalto University, Finland

Introduction

Dealing with historical geographic places (Southall et al., 2011) is important in museums, libraries, archives, and media companies, but challenging: 1) Historical places change in time. 2) It is difficult to understand the spatial and temporal context of the places. 3) Historical place names can often be seen only on historical maps. 4) Historical geographic data is scattered across multiple sources that can be incomplete and/or mutually conflicting. 5) To preserve semantic interoperability across Cultural Heritage (CH) datasets, there is a need to find out how the same place is represented in different repositories. 6) If a place is nowhere to be found—a situation quite common—there should be a mechanism to suggest and share new place concepts among the CH community.

Dataset	Original Source	Place type	Size	Description
Finnish Municipalities 1939–44	National Archives of Finland	Municipality	612	Finnish National Archives research project "Finland, prisoners of war and extraditions 1939–1955" produced a map application, from where the war time municipalities were obtained.
Karelian map names 1922–44	Jyrki Tiittanen / National Land Survey of Finland	village, house, etc.	34 938	Historical places in the Karelia region of Finland and Russia.
Finnish Spatio-Temporal Ontology	SeCo	Municipality	1 261	A spatio-temporal ontology of Finnish municipalities.
Finnish Geographic Names Registry	National Land Survey of Finland	61 place types	800 000	The place name dataset comprises natural and cultural names whose spelling has been checked by the Institute for the Languages of Finland.
The Getty Thesaurus of Geographic Names	J. Paul Getty Trust	1800 place types	2 156 896	TGN is a structured vocabulary containing names and other information about places. Names for a place may include names in the vernacular language, English, other languages, historical names, names and in natural order and inverted order. Among these names, one is flagged as the preferred name.
Senate atlas	National Archives of Finland	Map	414	Series of maps of Southern Finland drawn by the Russian Army topographic troops in the end of the 19th and the beginning of the 20th centuries in scale 1:21 000.
Karelian maps	National Land Survey of Finland	Map	47	The National Board of Survey and Topografikunta produced four-colour topographic maps in scale 1:100 000 during 1928–1951.

Table 1. Datasets connected to Hipla.fi

To tackle these challenges, we have developed a Linked Open Data brokering service model HIPLA for using and maintaining historical place gazetteers and maps based on distributed SPARQL endpoints. Using Linked Data technologies, HIPLA provides a common search interface to historical geographic data like place names with coordinates and historical maps. Contextual information, e.g. historical events or photographs related to a geographic location, is provided to help the user to gain a deeper understanding of the historical place. HIPLA also serves as a sustainable and evolving repository of historical places by implementing Dynamic Ontology Services for Evolving Ontologies (Hyvönen et al., 2015). Cultural Heritage organizations can connect their legacy cataloguing systems to HIPLA using a widget or an API in the same vain as in the ONKI ontology service (Tuominen et al., 2009).

The general HIPLA model is being implemented to create and manage a national level gazetteer and map service Hipla.fi. Hipla.fi is based on four Finnish datasets in SPARQL endpoints totalling some 840,000 geocoded places, on 450 historical maps from two atlas series aligned on modern maps, and on the Getty Thesaurus of Geographic Names (TGN) SPARQL endpoint in the US.

This paper first presents Hipla.fi's user groups (section 2) and the end-user interface (section 3), complementing the crowdsourcing view to the system (Hyvönen et al., 2015). Then the system architecture is outlined (section 4), and finally lessons learned are discussed (section 5). Hipla.fi is available at http://hipla.fi.

HIPLA user groups

The audiences of HIPLA are 1) collaborative geo-ontology developers, 2) cataloguers of historical content, 3) information searchers, and 4) application developers. For group 1 HIPLA facilitates a sustainable model for aggregating historical place names in shared data repositories as time goes by. For groups 2 and 3 HIPLA provides a combination of historical and contemporary maps, linked contextual data, and semantic federated search to find and understand historical places. User group 4 can utilize distributed SPARQL endpoints, URI resolving services, and an autocompletion text search widget.

Finding and understanding historical places in context

Federated search

Our first focus in developing Hipla.fi has been on modeling, storing, and searching Finnish place names in multiple SPARQL endpoints, and on displaying them on historical and contemporary maps. The datasets used are stored in separate RDF graphs, which makes it possible to offer dynamic selection of data sources for the user interface or external data consumers. Table 1 presents the datasets currently connected to Hipla.fi, most of them available on the Linked Data Finland platform[1] (Hyvönen et al., 2014).

Figure 1 depicts the Hipla.fi user interface. For finding, disambiguating, and examining historical places, there is an autocompletion search input field (a). Place names can be searched from multiple SPARQL endpoints at the same time based on the user's choice (checkboxes above (b)) with the following functionalities:

1. Hovering the cursor over the search results shows where the places are: the corresponding marker bounces on the map.

2. A click on a search result label opens the info window of the place, showing its context (c).

3. A click of the menu button on a result row (a) shows the place data in a Linked Data browser for investigating the data in detail.

Map-based multiple dataset browsing

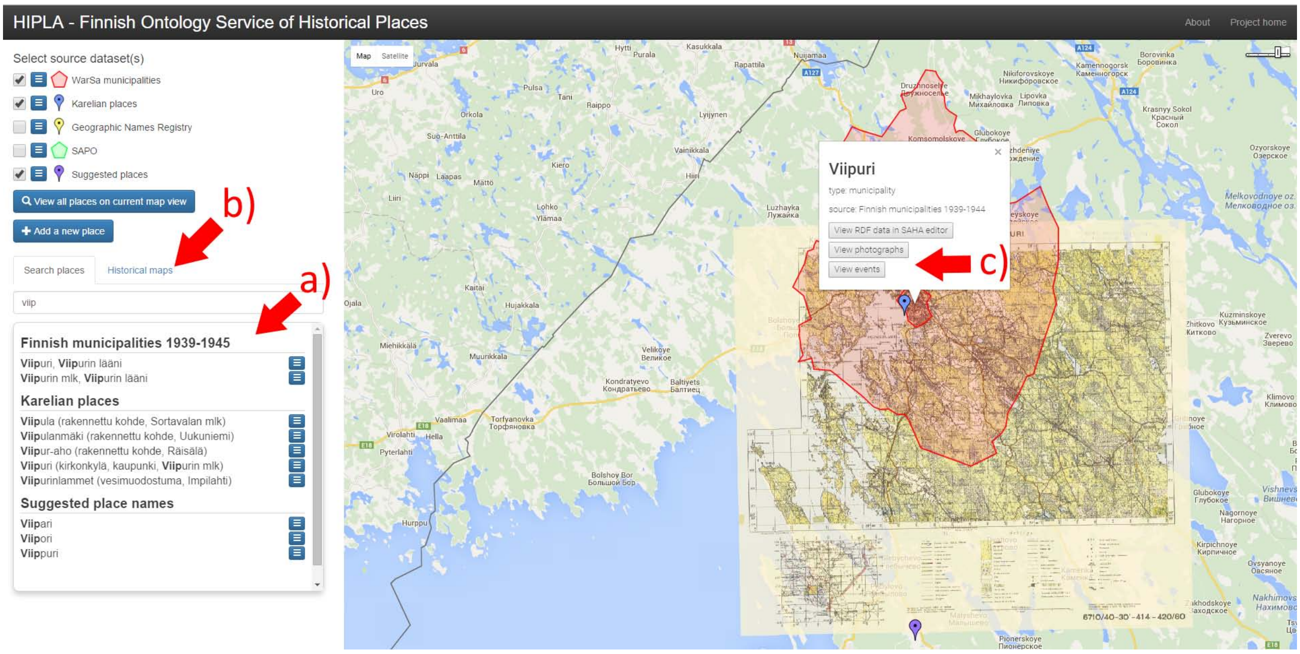
If the user does not know the name of the place, but has some idea where the place is located, she can pan and zoom the map view to the area. After this it's possible to use the "View all places on current map view" button. This way places from different datasets connected to Hipla.fi are rendered on the map, and the user can check if the place exists already in some of the datasets, and compare places in different gazetteers.

Fetch historical maps

The "Historical maps" tab (Figure 1 (b)) provides a list of old maps that intersect the current map view. The map images are fetched from the Hipla.fi's Map Warper service[2] and their metadata is queried with SPARQL from the map RDF graph of the HIPLA service. Each map has a checkbox for rendering the map on the main map view, a thumbnail image, information about map series, scale and type, and a link to view the map in Map Warper. All map series are visible by default, but with the series button it is possible to filter the maps by their series. Once one or more historical maps have been selected with the checkboxes, the opacity of the historical maps can be controlled with the slider that is located on the top right corner of the map. If the user pans or zooms the main map view, clicking the "Refresh map list" button updates the map list.

View contextual data

When the user selects a place, contextual data (Figure 1 (c)) is provided for connecting the place to other relevant data sources. This functionality is first piloted with the



HIPLA - Finnish Ontology Service of Historical Places
About
Project home
Map
Satellite
b)
a)
c)

Select source dataset(s)

☑ WarSa municipalities
☑ Karelian places
□ Geographic Names Registry
□ SAPO
☑ Suggested places

View all places on current map view
Add a new place
Search places
Historical maps

viip

Finnish municipalities 1939-1945

Viipuri, **Viip**urin lääni
Viipurin mlk, **Viip**urin lääni

Karelian places

Viipula (rakennettu kohde, Sortavalan mlk)
Viipulanmäki (rakennettu kohde, Uukuniemi)
Viipur-aho (rakennettu kohde, Räisälä)
Viipuri (kirkonkylä, kaupunki, **Viip**urin mlk)
Viipurinlammet (vesimuodostuma, Impilahti)

Suggested place names

Viipari
Viipori
Viippuri

Viipuri

type: municipality

source: Finnish municipalities 1939-1944

Figure 1. Hipla.fi user interface

spatial datasets of the WarSampo portal (Hyvönen et al., 2016), providing, e.g. 160 000 historical photos of the Second World War related to the places, and a timeline of historical events. In addition to this, the spatial perspective[3] of the WarSampo portal uses customized Hipla.fi user interface elements to visualize wartime places and their connections to other WarSampo datasets.

Extend with new gazetteers

The HIPLA model is adaptable to various geographic data models and both contemporary and historical gazetteers. The only requirement is that the gazetteer is published in a SPARQL endpoint. Because there is no standard for how to express the temporal extent of spatial data, the spatial dimension of gazetteer data can be utilized in the user interface (e.g. when disambiguating place names) by individual configurations.

System architecture

Figure 2 depicts the components of the HIPLA model. The Hipla.fi prototype is implemented using the Linked Data Finland platform (Hyvönen et al., 2014), based on Fuseki[4] with a Varnish[5] front end for serving the linked data. The end-user interface of Hipla.fi is a lightweight HTML5 single page map application, which provides access to multiple data sources with SPARQL queries and autocomplete search functionality using typeahead.js[6]. Embedded Google Maps view is used to visualize historical places. Hipla.fi's Map Warper is an instance of the open source Map Warper tool of the New York Public Library for georectifying old maps on top of modern ones.

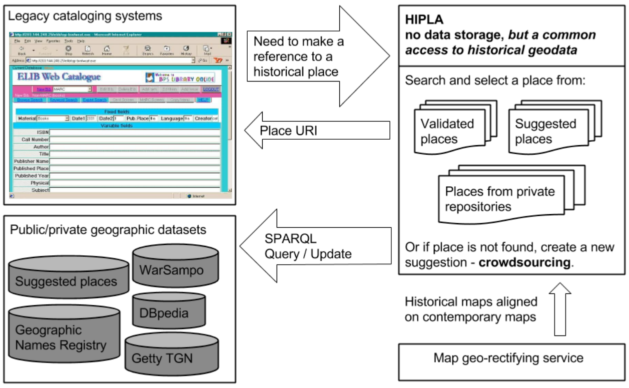

Legacy cataloging systems
ELIB Web Catalogue
Need to make a reference to a historical place
Place URI
HIPLA
no data storage, but a common access to historical geodata
Search and select a place from:
Validated places
Suggested places
Places from private repositories
Or if place is not found, create a new suggestion - crowdsourcing.
Public/private geographic datasets
Suggested places
WarSampo
DBpedia
Geographic Names Registry
Getty TGN
SPARQL Query / Update
Historical maps aligned on contemporary maps
Map geo-rectifying service

Figure 2. HIPLA system architecture

Related work and discussion

HIPLA is an ontology library service (d'Aguin and Noy, 2012) for historical places. Complementing traditional gazetteers, HIPLA not only publishes the data for humans but also for machines using the SPARQL endpoint API. In addition, historical maps and contextual linked data about the places are provided.

Thesauri of historical places, published as Linked Data, include the Getty TGN[7] of some 1.5 million records, 'Pelagios: Enable Linked Ancient Geodata In Open Systems'[8], and Pleiades[9]. Pelagios and Pleiades are based on crowdsourcing volunteers' work in ontology development. The novelty of HIPLA from a user interface viewpoint lays in the idea of combining multiple geographic data sources, offering a unified view for examining and comparing them. In addition, HIPLA makes it possible to crowdsource the creation of the gazetteer to cataloguers of Cultural Heritage content, as a side effect of their daily work, as discussed in Hyvönen et al., 2015.

The Historical Gazetteer of England's Place-names[10] is a service of over 4 million names that can be searched and viewed on modern maps as well as on historical ones. HIPLA has a similar local flavor focusing on places in Finland, but is based on Linked Open Data. OldMapsOnline[11] is a search engine for finding historical maps covering a given area. In contrast to the systems above, HIPLA includes a map service for aligning and viewing georectified historical maps, as in the New York Public Library's Chronology of Place gazetteer[12]. HIPLA also publishes the metadata of the historical maps as Linked Open Data and the dynamic and transparent selection of data sources makes it possible to understand the origins of the data.

Our research was supported by the Finnish Cultural Foundation and Wikidata Finland.

Bibliography

’Aquin, M. and Noy, N. F. (2012). Where to publish and find ontologies? A survey of ontology libraries. *Web Semantics: Science, Services and Agents on the World Wide Web*, **11:** 96–111.

Hyvönen, E., Tuominen, J., Alonen, M. and Mäkelä, E. (2014). Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. *The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers*. Springer–Verlag, pp. 226–30.

Hyvönen, E., Tuominen, J., Ikkala, E. and Mäkelä, E. (2015). Ontology services based on crowdsourcing: Case national gazetteer of historical places. *Proceedings of the ISWC 2015 Posters & Demonstrations Track*, CEUR Workshop Proceedings, Vol 1486. http://ceur-ws.org/Vol-1486/.

Hyvönen, E., Heino E., Leskinen, P., Ikkala, E., Koho M., Tamper M., Tuominen, J. and Mäkelä, E. (2016). Publishing Second World War History as Linked Data Events on the Semantic Web. *Proceedings of Digital Humanities 2016*, short papers, Kraków, Poland, July 2016.

Southall, H., Mostern, R. and Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, **5:** 127–45.

Tuominen, J., Frosterus, M., Viljanen, K. and Hyvönen, E. (2009). ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*. Springer–Verlag, pp. 768–80.

Notes

[1] http://www.ldf.fi
[2] http://mapwarper.onki.fi
[3] http://www.sotasampo.fi/en/places/
[4] https://jena.apache.org/documentation/serving_data/
[5] https://www.varnish-cache.org
[6] https://twitter.github.io/typeahead.js/
[7] http://www.getty.edu/research/tools/vocabularies/tgn/
[8] http://pelagios-project.blogspot.fi/p/about-pelagios.html
[9] http://pleiades.stoa.org
[10] http://www.placenames.org.uk
[11] http://www.oldmapsonline.org
[12] http://nypl.gazetteer.us

Wired!: Collaborative Teaching & Critical Digital Making In An Art History Classroom

Hannah L. Jacobs
hannahlj@gmail.com
Wired! Lab for digital art history & visual culture, Duke University, United States of America

The Wired! Lab for digital art history and visual culture at Duke University comprises a group of faculty, staff, and students engaged in applications of visualization methods to studies of material culture and art, architectural, and urban histories. Members of the lab collaborate to develop critical digital research employing 3D modeling, mapping, and database tools. Art historians and digital humanists work together to integrate both digital and art historical methodologies in lab courses and projects.

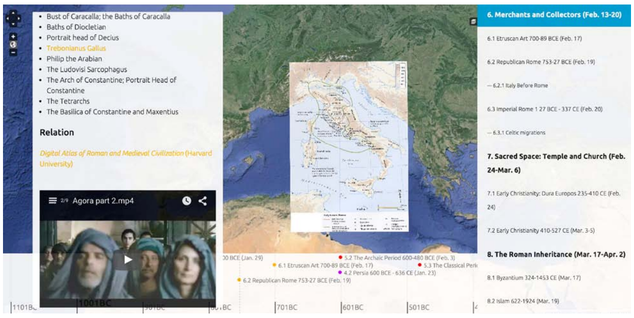
In the Wired! classroom's collaborative teaching model, a digital humanist takes on a significant role in both course planning and implementation. She works with instructors, graduate assistants, and librarians to redesign syllabi and assignments for preexisting departmental courses that incorporate not only digital tools but also critical methods. She then attends class meetings to familiarize herself with courses' art historical content; she delivers workshops on digital concepts and tools; and she works with instructors and students to establish project workflows, to troubleshoot technical issues, and to critique student work. For students, this kind of collaboration can provide opportunities to make intellectual connections across two modes of inquiry as they apply digital methodologies to art historical topics. For instructors, this collaboration can enrich pedagogical practice as digital methods present different possibilities for student engagement.

While some educators have focused on pedagogical challenges such as, "How does one teach students the digital tools to address a wide variety of projects without neglecting traditional discipline-specific issues of research formulation and data collection?", (Johanson and Sullivan et al., 2012) the Wired! Lab's digital pedagogy focuses on only the digital knowledge required for a specific topic. This approach ensures that students intentionally engage art historical content via digital methods, prioritizing quality of digital interventions over quantity while also addressing very practical issues of scalability within a disciplinary context.

In this presentation, I will examine two cases in which Wired! Lab instructors and a digital humanist collaborated to design and implement project-based undergraduate courses. These examples will demonstrate how the different teaching teams worked in tandem to create these learning experiences and will discuss benefits and challenges of these pedagogical collaborations. I will also situate the Wired! Lab's pedagogical work within the larger digital humanities and digital art history ecosystem.

Introduction to Art History

In Spring 2015, Professor Caroline Bruzelius implemented a team-based teaching approach for her Introduction to Art History course. Together, we worked with a graduate assistant and librarian to redesign the survey course and student projects. We all attended class meetings, we each taught aspects of the course, and we assessed student projects as a group. Combining our variant expertise, we created a course in which students employed a digital humanities approach to performing art historical critical analyses of spatial, temporal, and cultural relationships that influenced the movement of raw materials and cultural objects across ancient and medieval Western and Mediterranean societies. The digital tools we chose to use in the course were Neatline, a spatiotemporal exhibit builder, and Omeka, a collection management system in which Neatline operates. We implemented Neatline first for visualizing the syllabus and second for developing students' visual narratives concerning specific pre-modern art historical objects and materials.



- Bust of Caracalla; the Baths of Caracalla
- Baths of Diocletian
- Portrait head of Decius
- Trebonianus Gallus
- Philip the Arabian
- The Ludovisi Sarcophagus
- The Arch of Constantine; Portrait Head of Constantine
- The Tetrarchs
- The Basilica of Constantine and Maxentius

Relation

Digital Atlas of Roman and Medieval Civilization (Harvard University)

2/9 Agora part 2.mp4

6. Merchants and Collectors (Feb. 13-20)
6.1 Etruscan Art 700-89 BCE (Feb. 17)
6.2 Republican Rome 753-27 BCE (Feb. 19)
— 6.2.1 Italy Before Rome
6.3 Imperial Rome 1 27 BCE - 337 CE (Feb. 20)
— 6.3.1 Celtic migrations
7. Sacred Space: Temple and Church (Feb. 24-Mar. 6)
7.1 Early Christianity: Dura Europos 235-410 CE (Feb. 24)
7.2 Early Christianity 410-527 CE (Mar. 3-5)
8. The Roman Inheritance (Mar. 17-Apr. 2)
8.1 Byzantium 324-1453 CE (Mar. 17)
8.2 Islam 622-1924 (Mar. 19)

00 BCE (Jan. 29)
5.2 The Archaic Period 600-480 BCE (Feb. 3)
6.1 Etruscan Art 700-89 BCE (Feb. 17)
5.3 The Classical Peri
4.2 Persia 600 BCE - 636 CE (Jan. 23)
6.2 Republican Rome 753-27 BCE (Feb. 19)
1101BC
1001BC
701BC
601BC
501BC

Fig. 1. Neatline syllabus for Introduction to Art History

The interactive visual syllabus (Fig. 1) introduces students to the course narrative: its units and lectures are shown in time and space accompanied by contextual maps, specific geospatial points of reference, and other relevant multimedia including hyperlinks to important objects' museum pages, lecture slides, and supplementary videos. The interactive visual syllabus makes explicit temporal, spatial, and cultural relationships that effected the development of art practices across pre-modern societies. Presenting the syllabus in Neatline also familiarized students with Neatline's affordances and interface in preparation for creating their own Neatline projects, in which they used critical understandings of spatiotemporal narrative to develop cohesive art historical arguments concerning particular pre-modern objects, their making processes, political influences, and economic and environmental impacts.

Visualizing Venetian Art

Dr. Kristin Lanzoni's upper level course on early modern Venetian art also employed a collaborative teaching model. She and I worked together to develop a syllabus and project in which students studied course material through processes of visualization. Students spent the semester not only learning about Venetian art, history, and culture but also working together to model a Venetian palace no longer extant and to design an immersive visual narrative about the palace's political and cultural significance (Fig. 2). The students worked with tools ranging from Adobe Photoshop to SketchUp to Unity3D to visualize the palace.

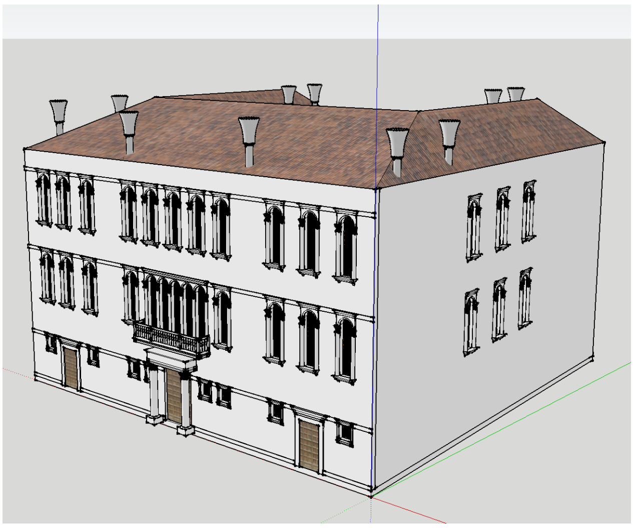


Fig. 2. Model of a Venetian palazzo created by students in Visualizing Venetian Art

As the semester progressed, Dr. Lanzoni and I worked closely with the students to troubleshoot a number of problems that arose as students strove to translate historical evidence into an historically informed 3D model. These issues stemmed from both primary sources, which give conflicting visual evidence for the palace's scale and appearance, and digital tools, which present challenges for modeling non-rectilinear structures and force compromises with regard to levels of detail. Students had to make joint decisions regarding model and narrative designs based on both historical research and the digital tools' affordances and limitations.

In both courses, students gained understandings of art historical topics through digital visualization processes. Wired! Lab teaching teams facilitate these types of learning experiences by combining their expertise in course design and implementation. While in the survey course, students were asked to create individual visual narrative projects, guided by an art historian, a graduate student, a librarian, and a digital humanist, the students in Visualizing Venetian Art worked together on a single topic from which their learning about early modern Venice radiated outward. While some digitally-inflected Wired! courses ask students to work individually, other courses model collaboration not only in the teaching but also in the learning. Overall, the majority of Wired! courses are not "Introduction to Digital Humanities" but rather art history courses redeveloped to integrate specific digital approaches that directly support course-specific content and disciplinary methods. The integration of a digital humanist in Wired! art history courses ensures that students' digital projects are informed not only by disciplinary knowledge but also by critical approaches to digital methodologies.

Bibliography

Johanson, C. and Sullivan E. et al. (2012). Teaching Digital Humanities through Digital Cultural Mapping. In Hirsh, B.D. (ed) *Digital Humanities Pedagogy*, Open Book Publishers, pp. 121-49.\

Comparison of Methods for the Identification of Main Characters in German Novels

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Wuerzburg, Germany

Isabella Reger
isabella.reger@uni-wuerzburg.de
University of Wuerzburg, Germany

Markus Krug
markus.krug@uni-wuerzburg.de
University of Wuerzburg, Germany

Lukas Weimer
lukas.weimer@stud-mail.uni-wuerzburg.de
University of Wuerzburg, Germany

Luisa Macharowsky
luisa.macharowsky@stud-mail.uni-wuerzburg.de
University of Wuerzburg, Germany

Frank Puppe
frank.puppe@uni-wuerzburg.de
University of Wuerzburg, Germany

Motivation

Digital literary studies have embraced social network analysis as a powerful tool to formalize and analyze social networks in literary texts (Elson et al., 2010b, Hettinger et al., 2015). Extracting networks automatically from texts is still a challenging task with the following steps: identification of all character references (which is not identical to named entity recognition), coreference resolution (CR) and a final step defining the amount of interaction between the characters, for example by the amount of verbal exchanges or the co-occurrence in a text segment. In the following we will discuss different ways to solve this task using an annotated corpus of German novels. One of the related problems is the definition of an evaluation metric which connects the computational problem to literary concepts like "main characters" and "character constellation". Our goal is to find the best way to capture the intuition behind these literary concepts in a formalized procedure. For this purpose we introduce a new way of evaluating automatically extracted networks. We make use of carefully created and revised summaries of German novels, provided by Kindler Literary Lexicon Online[1]. Besides, this work is to the best of our knowledge the first to compare different methods of creating and evaluating automatically extracted character networks.

Related Work

Social Network Analysis (SNA) is a well-established discipline, e.g. in the social sciences, which literary studies can apply for the analysis of character networks (Trilcke, 2013). Approaches to automatic extraction of SNs from literary text using NLP techniques have been manifold.

Most works start by identifying entities in the text and connect them via CR. Park et al. (Park et al., 2013) extract SNs based on proximity of names in the text and define a kernel function to distinguish protagonists from less important characters. Celikyilmaz et al. (Celikyilmaz et al., 2010) use an unsupervised actor-topic-model to create SNs from narratives. Elson et al. associate speakers with direct speech passages in novels (Elson et al., 2010a) and create SNs from the dialogues to validate literary hypotheses like whether the amount of dialogues is inversely proportional to the amount of characters that appear in the novel (Elson et al., 2010b).

Moreover, three end-to-end systems for the extraction and visualization of SNs from English literary texts already exist: PLOS (Waumans et al., 2015) works similarly to the approach by Elson et al. by creating networks from dialogue interactions. He et al. use their own speaker identification system to detect family connections between entities (He et al., 2013). SINNET by Agarwal et al. (Agarwal et al., 2013b) finds different types of directed events in a text and creates a directed SN from these events.

Data

This work is based on a corpus of 452 German novels from the TextGrid Digital Library[2]. Expert plot summaries from Kindlers Literary Lexicon Online are available for 215 of these novels. As the following experiments are partly based on direct speech, we analysed the novels with regard to the direct speech they contain. We selected 58 novels with the highest possible amount of direct speech for which there was also a summary on hand.

Those 58 novels have been split into tokens and sentences with OpenNLP[3], POS-tagged and lemmatized by the TreeTagger (Schmid, 1995), further processed by the RFTagger (Schmid and Laws, 2008) and the morphological tagger from MATE-Tools[4]. Additionally, we use the dependency parser by Bohnet (Bohnet and Kuhn, 2012) to analyze the sentence structure. Named Entity Recognition is done with the tool by Jannidis et al. (Jannidis et al., 2015) and the rule-based component by Krug et al. is used for CR. The detection of the speaker and the addressee for each direct speech passage is also part of the CR (Krug et al., 2015). In the summaries from Kindler, Named Entities and Coreferences have been manually labeled by two annotators.

Methods

We use four different methods to identify the most central characters in the novels and evaluate their quality by comparison with the characters occurring in the Kindler summaries.

The first method relies only on the frequencies of the characters in the text: the most central characters are those appearing most often in the novel (coreferences resolved). The second methods counts only those entities that have at least once been detected as speaker or addressee of direct speech. The other methods each construct a different type of social network and make use of SNA to find the most central characters. The first network is based on co-occurrences of characters in the same window of text: an edge between two characters exists if they are mentioned in the same paragraph and the weight of the edge is the number of paragraphs in which this is the case. The second network is created using the dialogue structure of the text. For each direct speech for which both speaker and addressee could be detected, an edge is drawn between those two. Longer dialogues consequently lead to higher edge weights between the participants. Thus, both network types are undirected and weighted. Examples for networks that were created with those methods are shown in Figure 1.

To identify the most central characters we use the weighted degree of each node (i.e. the sum of the weights of all edges incident to a node) in decreasing order. This metric is most intuitively interpretable with regard to the importance of characters in a fictional world.

In the following paragraph, we compare the rankings with the summaries and discuss possible sources of error and their influence on the results.

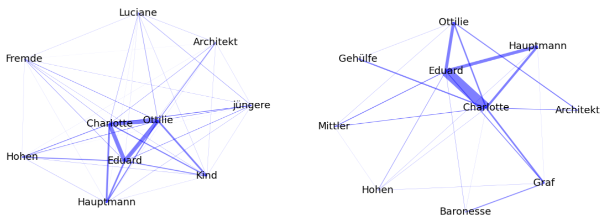


Figure 1: Automatically extracted SNs for Goethes: "Die Wahlverwandtschaften". The left picture shows the ten most connected characters when an interaction is created for a common appearance in a paragraph. The right picture shows the corresponding network when only direct speech is used as interactions.

Evaluation

Evaluating automatically extracted SNs is not a trivial task and there are no established practices. Elson et al. (Elson et al., 2010b) validate literary hypotheses, (Park et al., 2013) and (Waumans et al., 2015) analyze typical distributions that they expect of literary character networks. Agarwal et al. (Agarwal et al., 2013a) evaluate a machine-generated network of *Alice in Wonderland* against a manually conceived version by comparing typical SNA metrics like different centrality measures.

In this work, we want to compare the methods for identifying the most central characters as described in section 4. As a gold standard, we use the manually annotated Kindler summaries. The generated rankings for each novel, as well as the rankings from the summaries are first cleaned up so that only real names remain.

Our evaluation is based on the assumption that a summary contains all important characters. Since those summaries are carefully created and even revised by experts we propose that this assumption holds. For each summary, we create a ranking of the mentioned characters by [a] the number of occurrences (gold_count from here) and [b] the order of occurrence (gold_order from here). We relax the ranking assumption and only select the top 5 (top 10) figures from the summary rankings and compare them against the top 5 (top 10) characters in the automatically obtained rankings for the novels without respecting the particular ordering. If the name of a character from the gold standard is exactly found in an automatic ranking, there is a match. Table 1 shows the resulting correspondences with the two gold rankings, averaged over all 58 novels.

Results and Discussion

Table 1 displays first results for the identification of main characters in novels. Nevertheless, none of the methods yields very high scores for this kind of evaluation. Interestingly, the simpler approaches seem to be suited well for the task.

The low values can be explained by a variety of errors which can be grouped in three categories. Firstly, a character might not be among the top 10 of the relaxed ranking from Kindler. If automatic matches to lower positions in the ranking are allowed, the score in Table 2 can be reached.

Algorithm	DSN_Max	PN_Max	DSC_Max	Count_Max
gold_count	55.1%	56.6%	53.8%	57.3%
gold_order	58.0%	64.7%	55.1%	64.7%

Table 2: Accuracy of the matching, independent of the position in the automatic ranking

Algorithm	DSN_5	DSN_10	PN_5	PN_10	DSC_5	DSC_10	Count_5	Count_10
gold_count	40.5%	50.2%	39.3%	51.6%	38.9%	49.0%	40.1%	52.0%
gold_order	38.6%	45.1%	41.3%	48.6%	37.5%	45.3%	41.2%	48.5%

Algorithm	DSN_5	DSN_10	PN_5	PN_10	DSC_5	DSC_10	Count_5	Count_10
gold_count	40.5%	50.2%	39.3%	51.6%	38.9%	49.0%	40.1%	52.0%
gold_order	38.6%	45.1%	41.3%	48.6%	37.5%	45.3%	41.2%	48.5%

Table 1: Overview of the successfully matched entities between the two relaxed rankings from the summaries (gold_count, gold_order) and the generated relaxed rankings for the top 5 and the top 10 entities (DSN= Direct Speech Network; PN = Paragraph Network; DSC = Direct Speech Count; Count = simple frequency)

We can see that approximately 60% of the characters can now be matched unambiguously.

The highest percentage of errors is due to incorrectly resolved coreferences. Clusters of the same character that have not been merged during the CR do not only create redundant elements in the rankings, wrongly merged clusters also mean, that one character can never be matched correctly. If coreference errors are ignored, the results are as shown in table 3.

Algorithm	DSN_ Maxcr	PN_ Maxcr	DSC_ Maxcr	Count_ Maxcr
gold_ count	79.7%	81.2%	78.8%	81.2%
gold_order	58.6%	65.3%	55.6%	65.3%

Table 3: Accuracy of the matching, independent of the position in the automatic ranking, CR errors ignored

The third error type of originates from different spellings of the same name which make an unambiguous matching very difficult (e.g. "Amanzéi" vs. "Amanzei", "Lenore" vs. "Leonore"). Those kinds of errors are caused by different encodings, since the novels and the summaries originate from separate sources. Further reasons which render the matching more difficult or impossible respectively are missing or incorrectly detected Named Entities. The error analysis shows that future improvements are especially needed for the CR or procedures which avoid CR, since those have a better chance to succeed.

Conclusion

In this paper we showed work in progress to extract SNs from German novels. We compared four different approaches to the identification of central characters and evaluated against manually annotated summaries. Two presented methods rely on direct speech, the other methods can be applied to any novel. At least for this task, the more challenging approaches of determining speaker and addressee of direct speech and creating networks from the resulting interactions did score slightly lower than the simpler approaches. To improve the results, future work especially needs to be invested into the creation of a less error-prone CR system.

Bibliography

Agarwal, A., Kotalwar, A. and Rambow, O. (2013a). Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland, *Proceedings of the* 6th *International Joint Conference on Natural Language Processing (IJCNLP 2013),* Nagoya, Japan.

Agarwal, A. et al. (2013b). Sinnet: Social Interaction Network Extractor from Text, *Proceedings of the* 6th *International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japan.

Ardanuy, M. C. and Sporleder, C. (2014). Structure-based Clustering of Novels, *Proceedings of EACL 2014,* Gothenburg, Sweden.

Bohnet, B. and Kuhn, J. (2012). The Best of Both Worlds: a Graph-based Completion Model for Transition-based Parsers, *Proceedings of EACL 2012,* Avignon, France.

Celikyilmaz, A. et al. (2010). The Actor-Topic Model for Extracting Social Networks in Literary Narrative, *NIPS Workshop: Machine Learning for Social Computing*.

Elson, D. K. and McKeown, K. (2010a). Automatic Attribution of Quoted Speech in Literary Narrative, *Proceedings of AAAI 2010*, Atlanta, Georgia.

Elson, D. K., Dames, N. and McKeown, K. (2010b). Extracting Social Networks from Literary Fiction, *Proceedings of ACL 2010,* Uppsala, Sweden.

Gruzd, A. A. and Haythornthwaite, C. (2008). Automated Discovery and Analysis of Social Networks from Threaded Discussions, *Proceedings of INSNA 2008*, St. Pete Beach, Florida.

Hassan, A., Abu-Jbara, A. and Radev, D. (2012). Extracting Signed Social Networks from Text, *Workshop Proceedings of TextGraphs7 on Graph-based Methods for Natural Language Processing*, Jeju, Republic of Korea.

He, H., Barbosa, D. and Kondrak, G. (2013). Identification of Speakers in Novels, *Proceedings of ACL 2013,* Sofia, Bulgaria.

Hettinger, L. et al. (2015). Genre Classification on German Novels, *Proceedings of the 12th International Workshop on Text-based Information Retrieval,* Valencia, Spain.

Jannidis, F. et al. (2015). Automatische Erkennung von Figuren in deutschsprachigen Romanen, *Digital Humanities im deutschsprachigen Raum*, Graz, Austria.

Jing, H., Kambhatla, N. and Roukos, S. (2007). Extracting Social Networks and Biographical Facts from Conversational Speech Transcripts, *Proceedings of ACL 2007*, Prague, Czech Republic.

Krug, M. et al. (2015). Attribuierung direkter Reden in deutschen Romanen, *Digital Humanities im deutschsprachigen Raum 2016,* Leipzig, Germany, 2016.

Krug, M. et al. (2015). Rule-based Coreference Resolution in German Historic Novels, *Proceedings of the* 4th *Workshop on Computational Linguistics for Literature*, Denver, CO, USA, 2015.

Park, G. M. et al. (2013). Complex System Analysis of Social Networks Extracted from Literary Fictions, *International Journal of Machine Learning and Computing* **31**: 107-11.

Schmid, H. and Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Finegrained POS Tagging, *Proceedings of Coling 2008*, Manchester, UK. .

Schmid, H., Fitschen, A. and Heid, U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection, *Proceedings of LREC 2004,* Lisbon, Portugal.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German, *In Proceedings of the EACL SIGDAT Workshop 1995,* Dublin, Ireland.

Sutton, C. and McCallum, A. (2006). An Introduction to Conditional Random Fields for Relational Learning. In Getoor, L. and Taskar, B. (Eds.), *Introduction to Statistical Relational Learning*. Cambridge: MIT Press, pp. 93-128.

Trilcke, P. (2013). Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. In Philip Ajouri, P., Mellmann, K. and Rauen, C. (Eds.) , *Empirie in der Literaturwissenschaft,* Münster: Mentis, pp. 201-47.

Waumans, M. C., Nicodème, T. and Bersini, H. (2015). Topology Analysis of Social Networks Extracted from Literature, *PloS one* 10.6: e0126470.

Notes

[1] http://kll-online.de

[2] https://textgrid.de/digitale-bibliothek

[3] https://opennlp.apache.org/

[4] https://code.google.com/p/mate-tools/

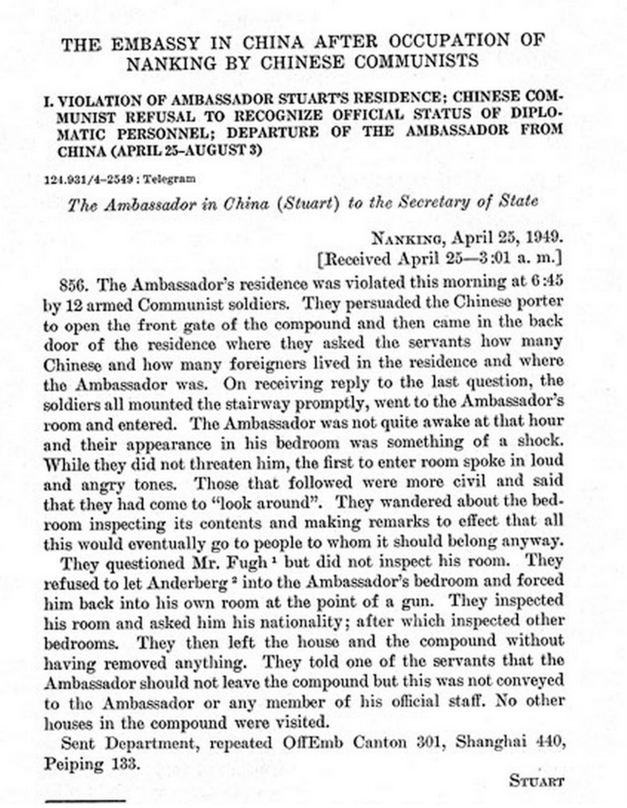
Diplomatic History by DataUnderstanding Cold War Foreign Policy Ideology using Networks and NLP

Eun Seo Jo
eunseo@stanford.edu
Stanford, United States of America

What is diplomatic culture or "rhetoric" and can we measure it?

This project is an attempt to understand quantitatively the language and structure of U.S. diplomacy as a bureaucratic institution through analysis of a large corpus of diplomatic papers. Since the "cultural turn" of History in the late twentieth century historians have produced cultural interpretations of American diplomacy that highlight gender and racial influences and justifications in diplomatic decision making but there have not been attempts to measure longue duree changes or to quantify them by a standardized measure.[1] This work hopes to fill this gap by introducing a new method of analyzing time-dependent bodies of text. I then apply these methods to a corpus of diplomatic papers to systematically chart changes in concepts and ideology detectable in diplomatic language.

The project has two aims. First, I am designing a method of measuring how a concept, as a fixed variable, evolves through a temporal corpus. Tools such as GloVe and Dynamic Topic Modeling are two existing approaches that can be used to understanding the linguistic shift and topic distributions over time.[2] I argue, however, that these tools are not adequate yet for time-sensitive tasks such as tracing concepts over time and attempt to design a new method optimized for this task. Second, mindful of the particular characteristics of this corpus as a set of diplomatic papers I want to apply the most appropriate methods of analysis and engage with the existing historiography on the Cold War to engage with claims historians have made about Cold War ideology. I am trying to answer: what is "Cold War rhetoric" in high level diplomacy? how did it evolve over the decades? and how did it become propagated within the diplomatic institution? I approach these questions from two methodological angles - networks and NLP.



THE EMBASSY IN CHINA AFTER OCCUPATION OF
NANKING BY CHINESE COMMUNISTS

I. VIOLATION OF AMBASSADOR STUART'S RESIDENCE; CHINESE COMMUNIST REFUSAL TO RECOGNIZE OFFICIAL STATUS OF DIPLOMATIC PERSONNEL; DEPARTURE OF THE AMBASSADOR FROM CHINA (APRIL 25–AUGUST 3)

124.931/4–2549 : Telegram

The Ambassador in China (Stuart) to the Secretary of State

NANKING, April 25, 1949.
[Received April 25—3:01 a. m.]

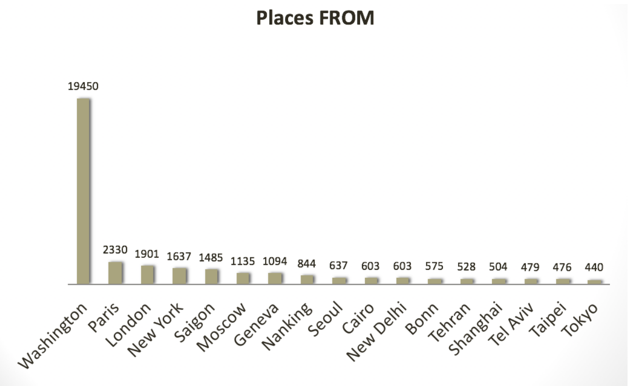
856. The Ambassador's residence was violated this morning at 6:45 by 12 armed Communist soldiers. They persuaded the Chinese porter to open the front gate of the compound and then came in the back door of the residence where they asked the servants how many Chinese and how many foreigners lived in the residence and where the Ambassador was. On receiving reply to the last question, the soldiers all mounted the stairway promptly, went to the Ambassador's room and entered. The Ambassador was not quite awake at that hour and their appearance in his bedroom was something of a shock. While they did not threaten him, the first to enter room spoke in loud and angry tones. Those that followed were more civil and said that they had come to "look around". They wandered about the bedroom inspecting its contents and making remarks to effect that all this would eventually go to people to whom it should belong anyway.

They questioned Mr. Fugh[1] but did not inspect his room. They refused to let Anderberg[2] into the Ambassador's bedroom and forced him back into his own room at the point of a gun. They inspected his room and asked him his nationality; after which inspected other bedrooms. They then left the house and the compound without having removed anything. They told one of the servants that the Ambassador should not leave the compound but this was not conveyed to the Ambassador or any member of his official staff. No other houses in the compound were visited.

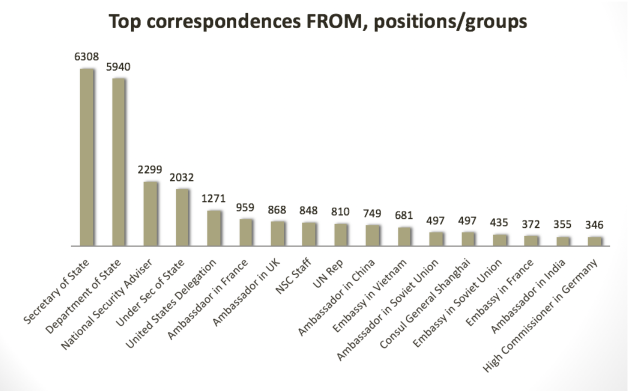
Sent Department, repeated OffEmb Canton 301, Shanghai 440, Peiping 133.

STUART

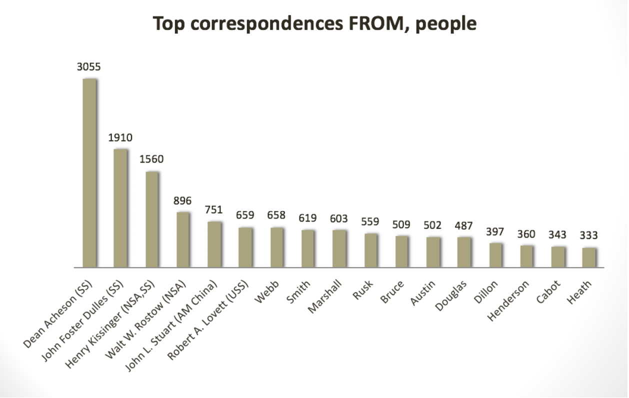
Image A: Example of FRUS document; Stuart as the Ambassador in China to the Secretary of State sent from Nanking on April 25, 1949



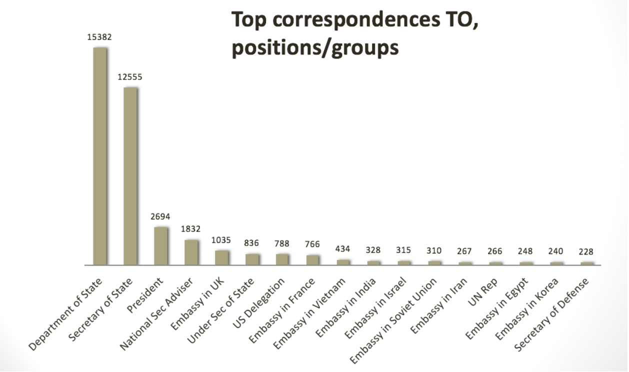
Graph 1A: Location of Correspondence Origin



Graph 1B: Top Correspondences sent from



Graph 1C: Persons that sent the most correspondences



Graph 1D: Top recipients of correspondences

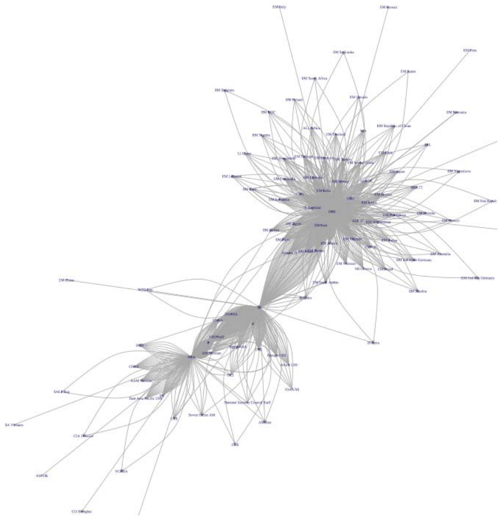
Dataset

The dataset for this analysis is the entire text corpus of the Foreign Relations of the United States (FRUS, 1861-1980), a collection of published declassified diplomatic papers. The FRUS collection is hand-curated by librarians at the Office of the Historian to be a representative and comprehensive sample of American diplomatic history and has been a dependable primary source base for historians and social scientists for decades. The document types include telegrams, airgrams, notes, and memoranda among others. Most documents have accompanying meta-data such as the name of the sender, name of the recipient, location of the sender, and date when applicable. So far, I have focused on a subset of this corpus, consisting of all papers from 1948 to 1980, to analyze the high Cold War era. This subset is also what is available in xml format online with hand-tagged meta-data and reliably edited text. About 92,000 documents were available for analysis. 3 A notable caveat of this work is that the corpus, while presumably a representative sample, is still a sample of non-random selection and could carry both deliberate and unintended biases of the librarians. The FRUS is a corpus that becomes publically accessible upon publication and has a fixed audience of social scholars. There is an inevitable feedback loop where the curators respond to the requests of the prime users of the collection such as adjusting the proportion of the most „useful" types of documents. For instance, as the Vietnam War is a highly contested field of scholarship curators may include a higher ratio of papers surrounding the Vietnam War thus distorting the overall representation of topics by exaggerating the war's significance. This possible limitation is something to keep in mind throughout the analysis of this corpus.

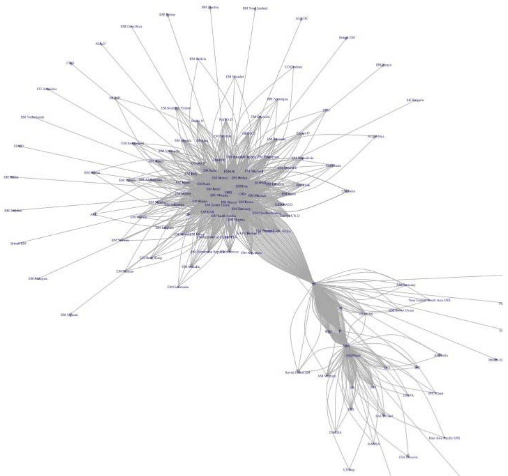
Analysis

Descriptive summaries of the meta-data from the corpus shows several clear distributive patterns. From the total set of all documents 42,000 were correspondences (have to and from agents) from which I parsed all available meta-data (name of sender, name of recipient, location of sender, date) and made inferences on missing data based on historical knowledge. The results show that the Department of State (DOS) as a bureaucratic institution is highly centralized around key actors. Not surprisingly, the prime location of correspondence origin is Washington and its overwhelming predominance indicates that the DOS correspondence system was used for sharing information from the center to the peripheries. Similarly, the top senders of the correspondents were concentrated in high administrative positions – the Secretary of the State (SS), Department of State (DOS) administrative center, and the National Security Advisor (NSA). The individuals that have the highest correspondence authorship are therefore those who have held SS or NSA positions such as Dean Acheson (SS), John Foster Dulles (SS), Henry Kissinger (SS, NSA), and Walt Rostow (NSA). The top recipients of correspondences are also SS and DOS confirming a much bilateral relation between central and peripheral offices.

I then mapped the network structure of correspondences to make the problem of bureacratic "culture" more concrete and visual. In this abstract, I have included images of correspondence networks from two discrete periods – when Kissinger served as Secretary of State from 1973-1977 and when Rostow served as National Security Adviser from 1966-1969. From the maps we can see the overall design of flow of information and transfer of knowledge based on the counts of correspondences and their directions. The maps confirm that indeed the bulk of the correspondence happens bilaterally between top administrative posts and peripheral agents. For instance, the DOS acts as the center point of correspondence for all embassies placed abroad and NSA as that for Washington based lower ranking administrative posts, such as the Under Secretaries of State. Further, two distinct communities emerge within the U.S. diplomatic institution. In both images, one can discern that the DOS and NSA act as distinct and separate focal points while the SS connects the two camps of correspondence and acts as a bridging agent of the two communities.

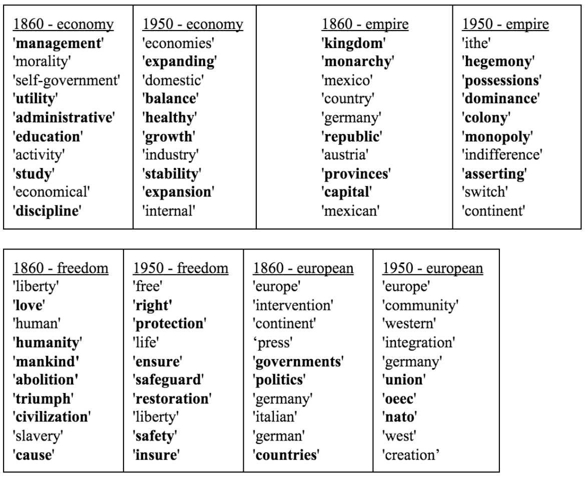


Graph 2A: Map of correspondence networks during the years Kissinger served as Secretary of State (1973-1977)



Graph 2B: Map of correspondence networks during the years Rostow served as National Security Advisor (1966-1969)

With this structural framework, I am using a combination of NLP methods to trace given "concepts" to see how they have changed in meaning over time. I identify concepts as individual terms, such as „liberty,“ or as a collection of related terms (topics). Word vectors are the most intuitive method of tracing changes in word meanings. Global vectors (GloVe) and other word vector models suppose a time stagnant corpus so I divided my corpus into decade-long chunks and worked with the assumption that language would not change in usage and meaning significantly enough to matter within ten-year time spans. My results comparing the nearing neighbors by Euclidean Distance of GloVe outputs of select concepts show there is a qualitative difference in conceptual meaning in the 1860s and 1950s. For instance, the concept „freedom“ in the nineteenth century was associated with more poetic and romanticized terms such as „triumph“ and „humanity“ whereas a century later it came to be linked with legal and defensive terms such as „right“ and „safeguard.“ Historians would contextualize this phenomenon in the American Civil War and the Civil Rights respectively. Then a question arises: Were diplomatic agents using terms that reflect the popularized lingo of their time or were their propagating it themselves? Who, in the DOS, initiated the usage of these concepts in such ways? Can we use the networks mapped above for to interpret this?



1860 - economy	1950 - economy	1860 - empire	1950 - empire
'management'	'economies'	**'kingdom'**	'ithe'
'morality'	**'expanding'**	**'monarchy'**	**'hegemony'**
'self-government'	'domestic'	'mexico'	**'possessions'**
'utility'	**'balance'**	'country'	**'dominance'**
'administrative'	**'healthy'**	'germany'	**'colony'**
'education'	**'growth'**	**'republic'**	**'monopoly'**
'activity'	'industry'	'austria'	'indifference'
'study'	**'stability'**	**'provinces'**	**'asserting'**
'economical'	**'expansion'**	**'capital'**	'switch'
'discipline'	'internal'	'mexican'	'continent'

1860 - freedom	1950 - freedom	1860 - european	1950 - european
'liberty'	'free'	'europe'	'europe'
'love'	**'right'**	'intervention'	'community'
'human'	**'protection'**	'continent'	'western'
'humanity'	'life'	‘press'	'integration'
'mankind'	**'ensure'**	**'governments'**	'germany'
'abolition'	**'safeguard'**	**'politics'**	**'union'**
'triumph'	**'restoration'**	'germany'	**'oeec'**
'civilization'	'liberty'	'italian'	**'nato'**
'slavery'	**'safety'**	'german'	'west'
'cause'	**'insure'**	**'countries'**	'creation’

Chart A: Sample of select words' change of top 10 GloVe neighbors from 1860s and 1950s ('economy', 'empire', 'freedom', 'european')

Discussion

This is a work in progress and there is still much work to be done in finding and developing new tools appropriate for time sensitive text data. Given history is a study of change and its significance, it is imperative that we do not assume a static distribution of words across time, eliminating many otherwise useful NLP tools. While Dynamic Topic Modeling considers time as a variable, it constructs a fixed number of topics based on a collection of words making it less favorable for corpora with minimal predictability and pattern as diplomatic papers. Unlike academic journals such as *Science* as Blei and Lafferty have applied their modelling on, diplomatic papers are much less consistent in topics. I have also discovered based on my experience of implementing these tools on the FRUS corpus that because in diplomatic papers certain topics predominate discussions at certain dates, I need to be aware of isolating these topics from purely semantic changes. For instance, in the 1940s, “communism” is closest neighbor based on GloVe results to “chinese” or “ccp” because of the Chinese Communist Revolution of 1949, which does not yield any surprising result about the meaning of “communism” in diplomacy.

Bibliography

1 See how Hoganson introduces a gendered interpretation of U.S. involvement in the Spanish-American and Philippine-American Wars. Hoganson, Kristin L. (2000). Fighting for American Manhood: How Gender Politics Provoked the Spanish-American and Philippine-American Wars (Yale Historical Publications Series).

2 **Pennington, J., Socher, R. and Manning, Ch. D.** (2014). GloVe: *Global Vectors for Word Representation*; David M. Blei and John D. Lafferty (2006) Dynamic Topic Models.

3 This dataset is available as manually labeled xml formatted files on GitHub (https://github.com/HistoryAtState/frus), which makes the corpus more reliable and meta-data accessible than the OCR scanned files of documents from 1861-1947.

Exploring The History Of The Qur’an Digitally

Tobias Jocham
jocham@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Michael Marx
marx@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Oliver Pohl
opohl@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Markus Schnöpf
schnoepf@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Initiated in 2007, the project Corpus Coranicum of the Berlin-Brandenburg Academy of Sciences and Humanities aims at building a comprehensive digital information system by providing access to relevant materials for the history of the Qur’an such as digitized versions of the oldest qur'anic manuscripts and their transliterations, comparisons of variant readings for each verse, texts from the environment of the Qur’an as well as providing commentaries for each sura, taking all the aforementioned elements into account. Both, manuscript evidence and variant readings, can be seen as the foundation for a future critical edition of the Qur'an. Following the German philological approaches to the history of the Qur'an before World War II such as the “Wissenschaft des Judentums” – a reform movement founded by Abraham Geiger (1810-1874) – and Gotthelf Bergsträßer's (1886-1933) “Korankomission” of the Bavarian Academy of Sciences, the Qur'an project in Potsdam is working on implementing a sustainable solution for exploring the history of the Qur’an, this time digitally.

This information system does not confine itself to the digital reproduction of the holy text but utilizes international standards like XML, Unicode and TEI to ensure the long-term readability and archivability of the conducted research, text analyses and editorial efforts. The print edition of the Qur'an produced in Cairo in 1924 is used as a reference for the documentation of the material for the textual history, since that print had a tremendous influence on following prints during the 20th century. following the analytical approach of Theodor Nöldeke (1836-1930), the project produces rich philological commentaries for each sura, exposing their chronological order and putting emphasis of the development of their literary forms across the 22 years of the prophet's proclamation.

Viewing the Qur’an as a text proclaimed in Arabia in Late Antiquity, the Corpus Coranicum project provides access to a collection of testimonies labeled as “Texte aus der Umwelt des Korans” (“Texts from the Environment of the Qur'an”). There, texts in Hebrew, Syriac, Greekt, Arabic, Ancient South Arabian, Ethiopian and others are being gathered, transcribed and translated, in order to highlight intersections and point out differences between and other documents from Late Antique culture, religions and traditions: Thus, the messages of the Qur'an can be viewed and understood in their respective contexts and in a new light.

Furthermore, the project gathers archeological evidence and conducts radiocarbon datings of qur’anic parchments in an ongoing German-French cooperation (2011-2014 Coranica, from 2015-2018 Paleocoran) to contribute substantially to the understanding of the Qur'an's history and the emergence of Islam. A joint goal of Corpus Coranicum and Paleocoran is to bring together all manuscripts that were originally kept in old Cairo and are now scattered around the world in a digital format for presenting them in their original form and order. On top of mere digitizations of the manuscripts, the Corpus Coranicum provides modernized transliterations of the original Arabic scripture. These transliterations are being shown in a self-developed font “Coranica” since other Arabic fonts like MS Typesetting or Amiri fail to display all the necessary characters occurring in the relevant texts of the project.

Next to commentaries, contextualization and analyzing manuscripts, the Corpus Coranicum project is building a corpus of variant readings on a word-level. Since the earliest time, the various readers of the Qur'an have recited the text in their own way. For each sura, each verse and word-coordinate in the Qur'an, the project compares the variant readings according to the written source with each other in order to show the varying tradition and interpretation of the holy text.

With the variant data accumulated so far the variances cannot only be analyzed on a word-to-word basis, but they can also be utilized to compute a general similarity measure between readers of the Qur'an by mapping the variant readings into a vector space which can be represented as a multi-dimensional matrix. For each word occurring in Qur'an, the Qur'an matrix is being assigned a dedicated row. Each variant reading of that word at this particular sura-verse-word-coordinate creates a column in that row, assigning the variant reading as its value. The same procedure is then applied to create matrices for each reader of the Qur'an. Whereas the Qur'an matrix can have multiple non-empty values in a row, a reader matrix can only have one: the corresponding coordinate of the variant reading the reader uses at that particular sura, vers and word position. Since the cutting angle between two vectors or matrices represents the similarity to each other, the Euclidian distance (see below) is being utilized to compute that similarity measure.

All the aforementioned branches of the Corpus Coranicum project, the history of the text, the manuscripts, analyses of variant readings, the literary and chronological commentary as well as the texts surrounding and having influenced the Qur'an are bundled together to develop a new perspective on the text. The website of the Corpus Coranicum goes beyond a traditional digital edition of the Qur'an and can better be described as a digital framework or digital information system for the Qur'an, as a variety of tools and different texts are present.

The project tries to pick up and go digitally beyond where German Qur'anic science tradition has left off. On top of the content created and functionalities implemented thus far, the projects aims at extending its range of features by providing internationalized versions of the website (English, French, Turkish) as well as integrating the Rafi Talmon (1948-2004, Arabist, University of Haifa) concordance to offer chrono-morphological and statistical analyses of the holy text.

The talk will give an overview about the current state of the project, will portray philological approaches and their technical applications as well as present results of the similarity computations mentioned above.

Bibliography

http://www.corpuscoranicum.de
http://cl.haifa.ac.il/projects/quran/index.shtml

Adding Value to a Research Infrastructure Through User-contributed ePublications

Catherine Emma Jones
catherine.jones@cvce.eu
CVCE, Luxembourg

Lars Wieneke
lars.wieneke@cvce.eu
CVCE, Luxembourg

Introduction

The so-called 'web 2.0 revolution' heralded in the middle of the last decade has so far neither replaced the author as the source of genuine creation nor turned conventional production processes in research and industry upside down. Rather than blurring the boundary between experts and amateurs, it questions the role of the institution as a source of authority by empowering individuals, small groups and communities (Wieneke 2010). Even today, institutions struggle to turn the productivity and impulses of their respective user communities into clear added value that benefits the institution and the community alike. In this paper we present and discuss our experiences in the development, implementation and management of user-created content built using the MyPublications tool on CVCE.eu. MyPublications enables users to create and publish their own enhanced publications using documents and resources on the European integration process available on CVCE.eu.

The CVCE's goal is to contribute to a deeper understanding of European integration by developing a dedicated digital research infrastructure. One component of the infrastructure is a series of digital collections of publications (ePublications) on themes and topics associated with the European integration process, as curated by a researcher or team of researchers. The ePublications are themselves aggregations of diverse research objects (resources), including researcher-written contextual articles, historical documents, press articles, photographs, interactive diagrams and timelines as well as other multimedia material, each with a set of descriptive, publishable metadata and a unique and persistent identifier.

State-of-the-art ePublication frameworks are built upon a set of underlying principles encapsulated by the presentation of research knowledge alongside mechanisms for describing, sharing, discovering, reusing and repurposing the scientific content (Bechhofer, Roure et al., 2010). The CVCE's publication model is no different. It encapsulates the following principles: (1) the ability to provide aggregations of content derived from many different published objects; (2) the provision of a unique and persistent identifier, ensuring sustainability of access; (3) the possibility of tracing the steps a researcher took to produce the ePublication; (4) the potential to reuse objects in a different context; (5) the ability to change the way objects are used by repurposing them; (6) the ability to reuse objects in compliance with IPR constraints.

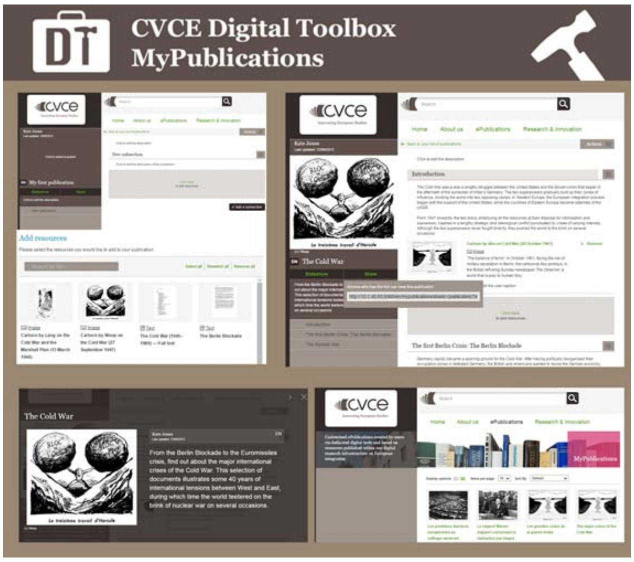
The current CVCE ePublication framework is expert-led, and research outputs are based on the centre's research questions and strategic topics. This approach leads to digital expert-curated ePublications based on themes and topics that are robust and reliable but — and this is both an advantage and a limitation — constrained by organisational priorities. On the other hand, the content itself can be combined and contextualised in different ways, both to highlight other perspectives and readings but also to address topics out of scope of the hosting institution. The challenge we face is therefore in the development of innovative tools and methods that offer new ways of reusing and repurposing historical objects by leveraging the potential of our user base to contribute knowledge themselves. This will in turn foster outputs that create a genuine surplus value for other users of the site by covering issues that we partially address or do not include at all.

The MyPublication tool empowers users to create tailor-made ePublications comprising resources in line with their personal needs, and encourages them to publish them on our website, thereby providing a plethora of different perspectives on the European integration process that ultimately enhance the value of our site for other users. The tool builds on the CVCE's multilingual ePublication model and infrastructure for research, teaching and learning activities in European integration studies.

The tool

MyPublications is the first instalment of the 'Digital Toolbox' at CVCE.eu. This suite of tools is designed to enable users to reuse or customise our resources for their own purposes. Using the MyPublications authoring tool, users can create ePublications that are personally curated collections of the many historical resources available on CVCE.eu alongside their own text, thoughts, ideas and critical analysis. The tool has been developed with the following workflow in mind: select, organise, structure, annotate, author (and edit), read, share and publish. The tool and its development are based on the CVCE's experience in building and maintaining a previous application called 'Albums', which was widely used by the teaching and learning community for assignments (e.g. 'build an album on the history of monetary union') or to document ongoing research (e.g. 'all resources related to the Rome Treaties').

Once a user is logged into the Digital Toolbox, they can browse collections of objects (referred to as 'resources' in the tool) or use the search facility to find and select relevant resources and add them to 'MyResources'. Users are then able to create a new publication using the MyPublication tool. They are asked to choose a cover picture, provide a short description and title and set the language. They then build a structure by creating subsections, add content to the different sections and write corresponding descriptions while being able to add, sort, organise and annotate the selected resources. Users can read the publication using a slideshow viewer, a simple interface that provides a sequential presentation of the narrative of an ePublication akin to a book. They can also share their publication with colleagues, peers and friends via a link. The publish/unpublish functionality enables users to make their publication accessible to the public in the MyPublications section of CVCE.eu (see figure 1). This should add new user-generated content to the research infrastructure.



From the Berlin Blockade to the Euromissiles crisis, find out about the major international crises of the Cold War. This selection of documents illustrates some 40 years of international tensions between West and East, during which time the world teetered on the brink of nuclear war on several occasions.

Figure 1: The MyPublications tool. Empty authoring interface (top left), MyResources section (middle left), slideshow view (bottom left), Share feature (top right), overview of publicly available MyPublications (bottom right).

The platform

The current CVCE data repository and ePublication framework uses the Alfresco document management system to manage, store and retrieve a wide variety of historical resources on the European integration process, such as interviews, treaties, legislation, photographs and cartoons, from an extensive array of archives and media outlets. The frontend and backend applications have been implemented in Liferay using HTML 5, CSS and JavaScript.

MyPublications was developed using the iterative design methodology Scrum, an AGILE development approach. This approach enabled us to involve our users in the development process on an ongoing basis, avoiding the need to predefine requirements for complex functionalities. The interface was designed with a focus on usability

(Shneiderman and Plaisant, 2005) and the aim of providing an easy-to-use, pragmatic environment. All the tools in the Digital Toolbox, especially MyPublications, were designed with a 'less-is-more' approach (Jones and Haklay, 2009) based on pragmatism and prioritising simplicity over complexity. As a benchmark we defined that it should be possible to explain how to use the tool with a two-minute help video. In-house users (historians, economists and political scientists) were engaged in testing, evaluating and providing feedback following each development cycle.

Conclusion

The simplicity of MyPublications enables a straightforward process of authoring and content creation for users with varying levels of digital literacy. Such facilities within research infrastructures also have the potential to increase accessibility to and reuse of existing objects to create new content. The tool provides a pragmatic solution for resources that have complex restrictions with respect to their licences for use. The content on our site is hosted by the CVCE, and full licences for permission to use the material on CVCE.eu have been acquired, so scholars, teachers and learners can avoid all the common problems associated with the use of weakly licensed material or unsustainable, transitory URLs. The option of publishing the publications in the CVCE research infrastructure increases the usability of the resources by enabling individual users, in various guises, to actively contribute to an international research infrastructure.

Following the release of the final tool in autumn 2015, we will reflect during our presentation at DH2016 on our experience of maintaining the MyPublications tool and user-created publications while particularly emphasising our lessons learned: how far have we succeeded in mobilising our user base to create and publish MyPublications? Is the incentive of being able to use licensed material, therefore avoiding copyright restrictions, strong enough, or is this not relevant for our users? What kind of content have our users created? Has the focus been on the final outcome (e.g. those publishing their own research) or has the process of creation been more relevant for them (e.g. those completing class assignments)? What tensions do we encounter between the institutional perspective and our user base regarding different perspectives and topics? And finally, what are the practical trials and tribulations of integrating user-created content in an institutional context?

Bibliography

Bechhofer, S., De Roure, D., Gamble, M., Goble, C. and Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. *The Future of the Web for Collaborative Science.*

Jones, C. E., Haklay, M., Griffiths, S. and Vaughan, L. (2009). A less is more approach to geovisualization–enhancing knowledge construction across multidisciplinary teams. *International Journal in Geographical Information Science*, **23**(8): 1007-93.

Shneiderman, S. B. and Plaisant, C. (2005). *Designing the user interface, 4th edition*. USA: Pearson Addison Wesley.

Wieneke, L. (2010). *An analysis of productive user contributions in digital media applications for museums and cultural heritage*. Weimar, Germany: Bauhaus-Universität Weimar. https://e-pub.uni-weimar.de/opus4/frontdoor/index/index/year/2010/docId/1442.

Implementing Canonical Text Services in the Croatiae Auctores Latini Digital Collection

Neven Jovanović
neven.jovanovic@ffzg.hr
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Alexander Simrell
arsimr16@g.holycross.edu
College of the Holy Cross, Worcester MA

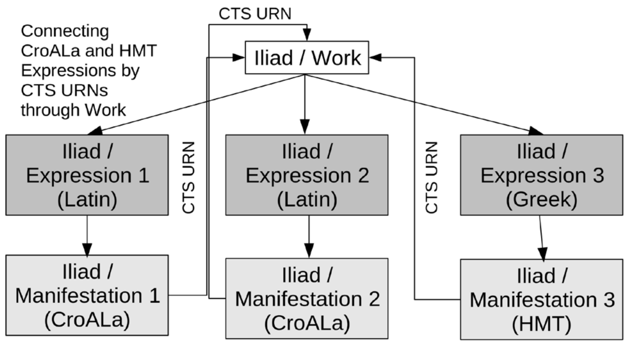
The Canonical Text Services (CTS) protocol (Blackwell and Smith, 2014) offers the scholarly community a way to use URNs for referring to two categories of propositional objects commonly called texts: to their ideal representations, **works**, and their specific realizations, **expressions** (International Working Group on FRBR and CIDOC CRM Harmonisation, 2015). CTS URNs point to complete texts and their subdivisions. CTS has a potential to transform scholarly practices. It supports the migration of our interpretations and knowledge from print to digital. It also forces us to reconsider what exactly we are doing when we refer or cite. It could, eventually, integrate into our referring and citing machine-driven comparisons across multiple versions of texts.

The CTS protocol is currently implemented in two projects: the *Homer Multitext* (Dué and Ebbott, 2015) and the *Perseus Digital Library* (Crane et al., 1987-). Both focus on texts which have traditionally been considered classical. Centuries, in some cases millennia, of appreciation and careful study have provided us with slightly different, but well-established citation schemes for such texts, and the main challenge to CTS up to now has been to reproduce these schemes. We put the protocol to a new test, by applying it to a non-canonical corpus of Latin texts published in the digital collection *Croatiae auctores Latini,* CroALa (Jovanović et al., 2009-).

CroALa collects and enables research of texts from a rich tradition of writing in Latin in Croatia. Latin was written through the Medieval and Early Modern periods up to modern times (our latest title is from 1984). The corpus includes a number of translations of Homeric poems into Latin, such as the partial one – an episode from the *Iliad* - by Janus Pannonius (1447), and a complete *Iliad* by Rajmund Kunić (1776). We wanted to connect

CroALa manifestations (digital editions) of these expressions to the *Iliad* as work, thus making possible a connection to manifestations of its other expressions, published elsewhere – in our case, to the Greek editions published by the *Homer Multitext*.

The process involved three stages: making the CroALa texts canonically referable through XML catalog records, validating and verifying the prepared editions, and establishing connections between editions prepared by different projects.


Connecting CroALa and HMT Expressions by CTS URNs through Work
CTS URN
Iliad / Work
Iliad / Expression 1 (Latin)
CTS URN
Iliad / Expression 2 (Latin)
CTS URN
Iliad / Expression 3 (Greek)
Iliad / Manifestation 1 (CroALa)
Iliad / Manifestation 2 (CroALa)
Iliad / Manifestation 3 (HMT)

Homer Multitext have produced URNs for each line of the *Iliad* as work; e. g. book 6, line 119 is described by urn:cts:greekLit:tlg0012.tlg001:6.119. A RDF triple connects this to URN of a line in the edition of the Venetus A codex (urn:cts:greekLit:tlg0012.tlg001.msA:6.119). Something similar is done for CroALa; you can see the work URN implied in the URN of a line our edition of Janus Pannonius' Latin translation (which is an expression fragment): urn:cts:greekLit:tlg0012.tlg001.croala-lat01:6.119. The same work URN is also implied by a line in the edition of Rajmund Kunić's complete Latin expression of the *Iliad*: urn:cts:greekLit:tlg0012.tlg001.croala-lat02:6.119.

The referencing happens through CTS Text Inventory (CTS TI), an XML catalog file. There Janus Pannonius' text is described by the following fragment:

```
<textgroup projid="greekLit:tlg0012">
 <groupname>Homer</groupname>
  <work projid="greekLit:tlg001">
    <title>Iliad</title> <translation xml:lang="lat"
    urn="urn:cts:greekLit:tlg0012.tlg001.croala-lat01"
    projid="greekLit:croala-lat01">
      <label xml:lang="lat">Diomedis et Glauci Congressus</label>
      <description>Jan Panonije [1447, Italia; Hungaria]:
      Diomedis et Glauci congressus, versio electronica,
      ed. Samuel Teleki
      Alexander Kovaczai</description>
      <online
      docname="/db/repository/greekLit/tlg0012/tlg001/
      tlg0012.tlg001.croala-lat01.xml">
        <citationMapping>
          <citation label="book"
          xpath="/div[@type="poesis-epica" and
          @n='?']" scope="/TEI/text/body">
            <citation label="line" xpath="/l[@n='?']" scope=
            "/TEI/text/body/div[@type='poesis-epica' and @n='?']"/>
          </citation>
        </citationMapping>
      </online>
    </translation>
  </work>
 (...)
```

The TEI XML guidelines allow multiple ways of marking up text structures (Schmidt, 2014). Therefore the most important sections in the fragment above are the XPaths which describe locations of individual books and linesin our editions. Books and lines can be encoded in a different way, represented by different XML elements, but through CTS URNs we are still able to connect the corresponding points, just as we refer to the same verse in the *Iliad*-as-work regardless of the fact that it is realised (printed or written) on different pages in different editions (Manifestation Product Types).

Our descriptions have also to be checked for correctness. Here the Homer Multitext project also can help; they have developed an excellent system for automated validation of editorial descriptions. We are adapting this to ensure that everything works in CroALa CTS, that no errors are introduced during the encoding process. Validation happens in a Virtual Machine which ensures that the entire process is replicable (Smith, 2015). But, since faithful replication of the process will only faithfully replicate systematic errors, a validation system was developed to assess our work in a different method, independent of how it was created.

The system first tokenizes all of the words. *Parsley*, a parsing machine for Latin morphology (Schmidt, 2015), checks that all tokens are valid Latin forms. Personal and geographic named entities do not parse automatically, so the system analyzes these separately. Named entities are checked for consistent markup and for compliance with our authority lists. The tokens that do not parse at either of these stages are analyzed by a researcher.

What has been validated has also to be verified as correct; our validation system ensures that all the tokens are acceptable Latin forms, but researchers have to ultimately decide whether the forms were correctly used.

Forms that are identified as being invalid are analyzed further for encoding errors, incorrect entries, problems with the parsing machine. One of recurring problems is that neo-Latin vocabulary is missing from the classical Latin dictionary used by the parser, so new words should be added to the lexicon. Even more numerous are neo-Latin forms that orthographically (sometimes morphologically too) violate classical Latin rules. Such forms have to be matched with the classical equivalents so that they can be accepted when the machine comes across these forms again. A similar approach would be required for all editions in which language differs markedly from the standard modern variant - e. g. for the Early Modern English as used in Shakespeare.

Once we have a text referenced by canonical URNs and tested as validated and verified CTS, we can serve the URNs and retrieve them from wherever we want. Connecting different editions - for example, linking Croatian Latin translations of Homer to editions of manuscripts prepared by the Homer Multitext project - is then a question of aligning the two sets of URNs. These aligned sets will enable us later, for example, to display in parallel the texts served behind each of them.

Though clear and simple in principle, the actual application of CTS to CroALa texts opened up a series of practical questions with certain theoretical implications. We mention only two.

First, a text and its translation are not always in a 1 : 1 relationship. A verse of the original can be rendered by verse and a half, or by a half verse, in the translation; a description ("Peleiades") can be glossed ("Achilles"). This had to be taken into account during the process of editorial verification. We had to introduce additional checks for translation alignments and establish a procedure for marking places where translation "shifts" equivalents forward or back in the textual structure (Latin equivalent of a Greek word appears elsewhere in the sentence, and therefore may appear in a different line).

Second, Croatian Latin translations of the *Iliad* are expressions of the Homeric work, but at the same time they themselves are of potential interest as authorial works, and they themselves exist in multiple manifestations (Kunić's translation was published in Rome 1776, Venice 1784, Vienna 1784, Firenze 1831 and 1838). To enable detailed scholarly study of translation as a work, the system has to take into account this additional level of multiplicity too: not only *Homer Multitext*, but also a *Kunić Multitext* (with the same underlying work).

Among the grand visions of digital humanities there is a dream of a world - or a space - in which different digital editions, carefully prepared, annotated and interpreted, mesh easily together, thus providing an especially rich and detailed groundwork for further annotations and interpretations. This space of interchangeability is today attained only rarely and with difficulty. The level of difficulty can be significantly lowered, as shown by CroALa's successful implementation of CTS and its automated validation and verification processes. A further step towards interchangeability will be wider acceptance of a digital canonical reference system such as CTS. For this to happen, a series of applications and "recipes" for specific usage cases is needed. We hope to have offered one such recipe here.

Bibliography

Blackwell, C. and Smith, N. (2014). *Homer Multitext.*Canonical Text Services protocol specification (accessed 4 March 2016).

Crane, G., Beaulieu, M.-C., Almas, B., Babeu, A. and Cerrato, L. (1987-). Perseus Digital Library (accessed 4 March 2016).

Dué, C. and Ebbott, M. (2015). The Homer Multitext project (accessed 4 March 2016).

International Working Group on FRBR and CIDOC CRM Harmonisation (2015). Definition of Object-Oriented FRBR (accessed 4 March 2016).

Jovanović, N., Haskell, Y., Lonza, N., Lučin, B., Marinova, E., Novaković, D. and Tunberg, T. O. (2009-). CroALa: Croatiae auctores Latini (accessed 4 March 2016).

Schmidt, D. (2014). Towards an Interoperable Digital Scholarly Edition. *Journal of the Text Encoding Initiative*(Issue 7) doi:10.4000/jtei.979. (accessed 4 March 2016).

Schmidt, H. (2015). goldibex/parsley-core *GitHub* (accessed 4 March 2016).

Smith, N. (2015). homermultitext/vm2015 *GitHub* (accessed 4 March 2016).

The Stanford Code Poetry Slam through Critical Code Studies

Melissa Kagen
mkagen@stanford.edu
Stanford University, United States of America

This paper reports on, analyzes, and contextualizes a project I have co-founded and run since 2013, the Stanford Code Poetry Slam (tinyurl.com/codepoetryslam). This is a series of international contests in which we solicit code poetry, whatever that means to our submitters, and then curate the best works. At each event, the best submissions are then “slammed” by human performers and (often) simultaneously by the computer programs that run them. The project explores the performative potential of computer languages, situates itself within the growing discipline of critical code studies, and has produced some fascinating work. In this short paper, I'll explain the project, analyze several of the poems as code and as poetry, and place the Stanford Code Poetry Slams in the context of recent conversations in critical code studies, particularly with reference to performance. How do humans perform code, and how does that differ from the way computers perform it? The connective aspect, I will argue, is based in the languages code poets use to write their works, each of which afford different performative possibilities. Through a close analysis of the linguistic choices our authors made, I will show in this brief presentation some of the commonalities between critical code studies and translation studies.

Many of our code poets have written and performed works that reference older poetic movements or forms, repurposing them for a digital medium. In Zak Kain's "Capsized," written in beautifully descriptive CSS, you can see a clear reference to imagistic works like William Carlos William's "The Red Wheelbarrow"; utilizing sparse but evocative descriptions, Kain's poem paints a harsh picture and simultaneously comes off as jokey and whimsical. This duality illustrates a fascinating, secretive aspect to code poetry, where the tone and content of the poem can imply one reading and the very language in which it's written (and the specific rules of that language) can inspire another reading entirely. In this case, the surrogate performer did a great job of presenting both aspects of the work, by starting sadly and then growing increasingly over-dramatic.

The code poem that won CPS 1.1 also played with performance to get across a profound point, in this case crossing seamlessly between aspects of performativity in digital, theatrical, religious, and social media realms. In "21st Century Prophecies," Hunter Bacot wrote a poem that calls the most recent tweet from seven "Prophets" (famous twitter users with huge followings) and strings those tweets together into a list of "virtues." In Keshav Dimri's performance, each line was intoned with the solemnity of a sermon ("Let KingJames be added to the list of virtues!") and the resulting poem (the 7 most recent tweets) was spoken like a biblical verse. "21st Century Prophecies" references movements like bricolage and found poetry by rearranging already extant text in new ways, and it gestures towards "Curation as Creation" and ideas found in the Digital Humanities Manifesto 2.0.

Other kinds of code poems make art out of the strenuous constraints imposed by coding languages, if one's goal is to write a text that actually compiles. Constrained texts reference much older poetic forms, like the sonnet, which require following a complicated set of rules about syllabic stresses and end rhymes. Many Perl poets write these kinds of texts. One of our best examples was Mike Widner's "A Pythonic Lament," which prints out "Alas! Alas!" when run.

Finally, in linguistic double coding, a sentence is readable in multiple languages at once. "Jean put dire comment on tape" reads in English (albeit a little ungrammatically) and, in translation from French, says "Jean is able to say how one types." Poems that are readable to humans and readable to computers perform a kind of cyborg double coding, and the ramifications of this possibility for translation studies are one of the themes the CPS series explores. Julian Bliss' "Polymorphism," the winning poem from CPS 2.0, took this idea to an incredible extreme, as he created a piece of text that, when compiled in multiple languages, produces a different poem in each. Moreover, each output poem parodies a clichéd English-language poem.

This presentation will analyze these and other works of code poetry we've slammed, showing how the languages in which they were written greatly affect their performative potential and demonstrate the performative nature of translation.

Trading Zones of Digital History

Max Kemman
max.kemman@uni.lu
University of Luxembourg, Luxembourg

As a subfield of the wider Digital Humanitiesdigital humanities, Digitaldigital hHistory is concerned with the incorporation of digital methods in historical research practices. Digital hHistory thus aims to do historical research using methods, concepts, or tools from other disciplines, making it a form of **methodological interdisciplinarity** (Klein, 2014){Formatting Citation}. However, how this interdisciplinarity affects the practices of Historyhistorians, on the methodological and the epistemological levels, remains underexplored. The PhD research presented in this paper aims to address this question by investigating the interdisciplinary interactions in which historians take part.

Three forms of interaction are of interest for this research, which are not necessarily an exhaustive list of digital history interactions. These forms are not mutually exclusive, but occur interchangeably and simultaneously, or one form could lead to another:

1. Digital history as collaboration with, among others, the computer science discipline.
2. Digital history as end-users of tools.
3. Digital history as building tools independently.

In order to look into such interactions, this PhD research will employ Galison's concept of **trading zones**, described as "an arena in which radically different activities could be locally, but not globally, coordinated" (Galison, 1996, p. 119).

When different groups interact with one another over a period of time in a trading zone, it is likely that the two groups will influence one another through **acculturation**: "the process by which the beliefs and practices of one community diffuse across the boundaries of another and subsequently alter the second community's practices and interpretations" (Barley et al., 1988). At the community level, acculturation involves changes of social structures, institutions, and cultural practices. At the individual level, it involves the behavioural repertoire of a person. By studying the acculturation of practitioners of digital history, as individuals and in groups, we may get a view of the types of trading zones and how these change over time. To model the different types of trading zones, we use three dimensions based on research by Berry (1997, 2005) and Collins et al. (2007):

1. Contact & Participation, i.e., how the two groups meet.

2. Cultural Maintenance (from homogeneous to heterogeneous), i.e., how the two groups define themselves and to what extent they aim to maintain their identity. On this scale, more homogeneous means the two groups become more alike to form a single group, while more heterogeneous means they remain two distinct groups.

3. Coercion (from collaborative to coercive), i.e., what the power relations in the trading zone are. On this scale, more collaborative means the two groups are both acting out of free will, while more coercive means one group is imposing practices upon the other.

The concept of trading zones has been used before to describe the digital humanities field. McCarty (2005) argues that humanities computing should rather be seen as a third space, neither belonging to one group nor the other, rendering it no longer a trading zone. However, in the terminology of Collins et al. (2007), this would constitute a collaborative-homogeneous trading zone, termed an inter-language. Svensson (2011, 2012a,b) suggests digital humanities is a collaborative-heterogeneous, termed fractioned, trading zone; a meeting place of two groups. Klein (2014) also describes digital humanities as a fractioned trading zone, and, like Svensson, emphasises that this may lead to a shared language, or jargon, between the different communities. Hunter (2014), without employing the concept of trading zones, describes digital humanities as a bridge or translation between two cultures, which we can describe as a collaborative-heterogeneous trading zone, termed interactional expertise. Rieder and Röhle (2012) use the concept to argue however that not the language should be central, but the interactions on the level of methodology, where not the terminology but the method itself is negotiated. In contrast to these authors, Mounier (2015) contends that there is a coercive political dimension underlying the field, which in the terminology of Collins et al. would suggest that digital humanities constitutes a coercive-heterogeneous, termed enforced, trading zone. This is not to say that this is how digital humanities will always be, but Mounier argues this should be better understood before we can move further and perhaps diffuse new digital methods into the wider humanities.

However, what is striking about these discussions of digital humanities as trading zones is that very little research into the **local** practices has been done, with the exception of Hunter (2014) who does not actually employ the concept of trading zones. Instead, digital humanities is discussed as a global phenomenon; this is in contrast with the original use of the concept by Galison as described above. This paper aims to reintroduce the concept of trading zones to describe local phenomena of digital humanities.

To this end, this paper investigates local manifestations of trading zones in digital history using the three dimensions described above. The analysis focuses on the first form of interaction described above, collaboration, and is based on interviews with practitioners of digital history, i.e., historians, computer scientists, and other collaborators, focusing on the diverse aspects of interdisciplinary collaboration. The interviews cover five distinct subjects, which together give an insight into the trading zones from each interviewee's perspective.

The first subject is that of **boundary work**, concerning how practitioners characterise their own discipline. Moreover, it is of interest how practitioners characterise the other disciplines in the collaboration. In previous research involving students of journalism, a lack of understanding what computer science is appeared to result in disinterest and performance anxiety (Cook, 2015). Furthermore, this subject covers the extent to which the practitioners aim to have their digital history research meet their discipline's values. This subject thus not only works towards the Cultural Maintenance dimension, but also already gives hints towards the Coercion dimension regarding how interested practitioners are in the collaboration with the other discipline.

The second subject is the **practice of research**, the research activities. This concerns a description of their research, both within and outside the trading zone, and the tools potentially used at different steps in their process.

The third subject concerns their **incentives** for practising digital history. In previous research on the collaboration between earth scientists and computer scientists, it was found that the two groups had different incentives for the collaboration (Weedman, 1998). This difference introduced difficulties for the collaboration and impacted the understanding of the other discipline. This subject thus works towards the Coercion dimension.

The fourth subject concerns the **organisation** of the collaboration. This concerns how often the groups meet, and where they are located, e.g., is it a collaboration between different departments at different places (or in different countries), or a sharing of office space. This subject thus works towards the Contact & Participation dimension.

Finally, the fifth subject concerns the **epistemological positions**. A criticism in the digital humanities debate is the incorporation of different epistemological positions such as positivism or objectivism in humanities scholarship (Drucker, 2011, 2013). It is therefore of interest whether practitioners in the trading zone (unconsciously) shift their epistemological position. A first question concerns their practice of reasoning; do they reason in a research question-driven deductive way, a more data-driven inductive way, or an abductive way to try to discover patterns (Dixon, 2012)? Other questions are related to epistemological positions. Roth and Roychoudhury (1994) developed a short qualitative questionnaire which allows to describe interviewees as more objectivist or more constructivist. Thus, this subject provides insight into the acculturation of practitioners, and works towards the Cultural Maintenance and Coercion dimensions on the epistemological level.

This paper will present preliminary findings of interviews held for this PhD research, focusing on the aspects of **incentives** and **organisation**. We will present a preliminary taxonomy of collaborations on the Contact & Participation dimension, and describe several digital history interactions on the Cultural Maintenance and Coercion dimensions. With these results, we aim to gain a better understanding of how digital history works as an interdisciplinary interaction, and how this impacts the practices of the involved groups and individuals.

Bibliography

Barley, S. R., Gordon, W. M. and Gash, D. C. (1988). Cultures of Culture: Academics, Practitioners and the Pragmatics of Normative Control. *Administrative Science Quarterly*, **33**(1): 24–60.

Berry, J. W. (1997). Immigration, Acculturation and Adaptation. *Applied Psychology*, **46**(1): 5–34. doi:10.1111/j.1464-0597.1997.tb01087.x.

Berry, J. W. (2005). Acculturation: Living successfully in two cultures. *International Journal of Intercultural Relations*, **29**: 697–712. doi:10.1016/j.ijintrel.2005.07.013.

Collins, H., Evans, R. and Gorman, M. (2007). Trading zones and interactional expertise. *Studies in History and Philosophy of Science Part A*, **38**(4): 657–66. doi:10.1016/j.shpsa.2007.09.003.

Cook, L. (2015). *Why Journalism Students Don't Learn CS*. https://source.opennews.org/en-US/learning/journalism-students-and-cs/ (accessed 27 October 2015).

Dixon, D. (2012). Analysis tool or research methodology? Is there an epistemology for patterns?. In Berry, D. M. (ed), *Understanding Digital Humanities*. Palgrave Macmillan http://eprints.uwe.ac.uk/16572/.

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, **5**(1): 1–21.

Drucker, J. (2013). Performative Materiality and Theoretical Approaches to Interface. *DHQ: Digital Humanities Quarterly*, **7**(1). http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html.

Galison, P. (1996). Computer simulations and the trading zone. *The Disunity of Science: Boundaries, Contexts, And Power*. Stanford University Press, pp. 118–57.

Hunter, A. (2014). Digital humanities as third culture. *MedieKultur: Journal of Media and Communication Research*, **30**(57): 18–33.

Klein, J. T. (2014). *Interdisciplining Digital Humanities: Boundary Work in an Emerging Field*. online. University of Michigan Press. http://hdl.handle.net/2027/spo.12869322.0001.001.

McCarty, W. (2005). *Humanities Computing*. Palgrave Macmillan.

Mounier, P. (2015). Une « utopie politique » pour les humanités numériques ?. *Socio*, **4**: 97–112. doi:10.4000/socio.1338.

Rieder, B. and Röhle, T. (2012). Digital Methods: Five Challenges. In Berry, D. (ed), *Understanding Digital Humanities*. Palgrave Macmillan, pp. 67–84.

Roth, W.-M. and Roychoudhury, A. (1994). Physics students' epistemologies and views about knowing and learning. *Journal of Research in Science Teaching*, **31**(1): 5–30. doi:10.1002/tea.3660310104.

Svensson, P. (2011). The digital humanities as a humanities project. *Arts and Humanities in Higher Education*, **11**(1-2): 42–60. doi:10.1177/1474022211427367.

Svensson, P. (2012a). Beyond the Big Tent. In Gold, M. K. (ed), *Debates in the Digital Humanities*. online. University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/22.

Svensson, P. (2012b). Envisioning the Digital Humanities. *DHQ: Digital Humanities Quarterly*, **6**(1). http://www.digitalhumanities.org/dhq/vol/6/1/000112/000112.html#.

Weedman, J. (1998). The Structure of Incentive: Design and Client Roles in Application-Oriented Research. *Science, Technology & Human Values*, **23**(3): 315–45. doi:10.1177/016224399802300303.

Converting the Liddell Scott Greek-English Lexicon into Linked Open Data using lemon

Fahad Khan
fahad.khan@ilc.cnr.it
Istituto di Linguistica Computazionale "A. Zampolli", Italy

Francesca Frontini
francesca.frontini@ilc.cnr.it
Istituto di Linguistica Computazionale "A. Zampolli", Italy

Federico Boschetti
federico.boschetti@ilc.cnr.it
Istituto di Linguistica Computazionale "A. Zampolli", Italy

Monica Monachini
monica.monachini@ilc.cnr.it
Istituto di Linguistica Computazionale "A. Zampolli", Italy

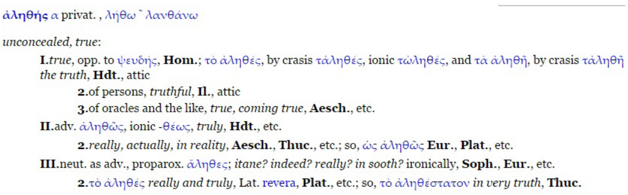
Introduction

The emergence and growing popularity of Linked Open Data (LOD) offers researchers a new range of possibilities when it comes to publishing datasets online (Hyvönen 2012, Oomen et al 2012); indeed not only does the success of LOD greatly facilitate the process of making scholarly data accessible and to a wider community but it also permits the enrichment of individual datasets by linking them to the other datasets available on the so called Linked Open Data Cloud. The advantages of Linked Open Data for teachers, academics and students in the humanities are obvious and are indeed manifold. However there is currently a paucity of linked open datasets in fields such as philology and literary studies, and in particular of datasets that deal with classical languages such as ancient Greek, Sanskrit, and Latin. This seems strange given the rich abundance of surviving works, of both a religious and secular character, that exist in those languages. A salient consideration here relates to the fact that even when such works have been digitised and made available in a format such as TEI-XML, a format which renders the structure and content of such texts more amenable to computer processing, the conversion of these resources into the Resource Data Framework (RDF), the standardised data model that underpins the Semantic Web, is not always straightforward.

In this article we describe ongoing work in the conversion of an important 19th century Ancient Greek resource the Liddell-Scott-Jones Lexicon, into RDF, part of a wider program of work that has been recently initiated at CNR-ILC in converting historical lexicons in languages such as Greek, Latin and Arabic into Linked Open Data.

Background

The Liddell-Scott-Jones lexicon (LSJ), or to give it its original title "A Greek-English lexicon", is a bilingual ancient Greek-English dictionary which since its first edition was published in the mid-nineteenth century has come to be regarded as amongst the most authoritative of modern day lexicographic resources dealing with the ancient Greek language, indeed it has the reputation of a standard in the field. As a result of its popularity the LSJ has been made available in a number of different versions differing in terms of the number of entries and the amount of data which they contain. For the work described in this paper we are using an abridged version of the LSJ which was originally published as "An Intermediate Greek-English lexicon," but which is more colloquially known as the "Middle Liddell" (ML). Fig. 1 shows the lexical entry for the adjective *ἀληθής* (alēthēs) in the ML. Entries in the ML are structured into different nested (sub-)senses, and each of these senses contain references (usually just the name of an author) attesting the use of the word as described in the sense's gloss.



ἀληθής α privat. , λήθω ˜ λανθάνω

unconcealed, true:

I.*true*, opp. to ψευδής, **Hom.**; τὸ ἀληθές, by crasis τἀληθές, ionic τὡληθές, and τὰ ἀληθῆ, by crasis τἀληθῆ *the truth*, **Hdt.**, attic

2.of persons, *truthful*, **Il.**, attic

3.of oracles and the like, *true*, *coming true*, **Aesch.**, etc.

II.adv. ἀληθῶς, ionic -θέως, *truly*, **Hdt.**, etc.

2.*really, actually, in reality*, **Aesch.**, **Thuc.**, etc.; so, ὡς ἀληθῶς **Eur.**, **Plat.**, etc.

III.neut. as adv., proparox. ἄληθες; *itane? indeed? really? in sooth?* ironically, **Soph.**, **Eur.**, etc.

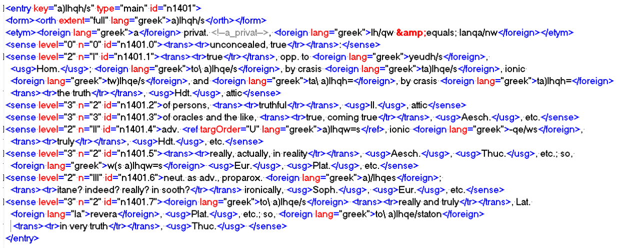
2.τὸ ἀληθές *really and truly*, Lat. revera, **Plat.**, etc.; so, τὸ ἀληθέστατον *in very truth*, **Thuc.**

Fig 1. An entry from the "Middle Liddell"

There were a number of motivations for choosing the LSJ as a starting point of our work into converting legacy lexical resources into linked data: aside of course from the question of its historical importance and continuing influence in the field of philology. Firstly we felt that given the lack of ancient Greek lexical resources in linked open data -- at the time of writing the Linguistic Linked Open Data cloud (Chiarcos et al 2011), that part of the LOD cloud that deals with linguistic data, contains no ancient Greek datasets -- there was an obvious necessity to ensure a presence on the cloud for a language that is absolutely foundational to the history of Western civilisation. Additionally there was also the challenge of converting a legacy resource like the LSJ which in its published form, and even in an abridged version like the ML, manages to condense a significant amount of lexical information in a relatively short amount of text, into linked data. In order to represent this information in the RDF model, and to stay close to the spirit of the Linked Open Data movement, a lot of what was implicit in the original text had to be teased out and rendered explicit.

Finally, one very important practical reason for choosing the ML was the fact that the conversion of the ML into XML using the TEI dictionary guidelines had already been carried out and made freely available under a creative commons license by the Perseus project (Crane et al 2013). This obviously saved us the trouble of digitizing the text ourselves and meant that we could work from a source file that was already annotated for lexical entries, senses, translations, etc.

In Fig 2 below we present the TEI-XML encoding of the ML entry for ἀληθής from the Perseus XML version of the ML which we used as our source dataset.



```
<entry key="a)lhqh/s" type="main" id="n1401">
<form><orth extent="full" lang="greek">a)lhqh/s</orth></form>
<etym><foreign lang="greek">a</foreign> privat. <!—a_privat—>, <foreign lang="greek">lh/qw &amp;equals; lanqa/nw</foreign></etym>
<sense level="0" n="0" id="n1401.0"><trans><tr>unconcealed, true</tr></trans>:</sense>
<sense level="2" n="I" id="n1401.1"><trans><tr>true</tr></trans>, opp. to <foreign lang="greek">yeudh/s</foreign>,
  <usg>Hom.</usg>; <foreign lang="greek">to\ a)lhqe/s</foreign>, by crasis <foreign lang="greek">ta)lhqe/s</foreign>, ionic
  <foreign lang="greek">tw)lhqe/s</foreign>, and <foreign lang="greek">ta\ a)lhqh=</foreign>, by crasis <foreign lang="greek">ta)lhqh=</foreign>
  <trans><tr>the truth</tr></trans>, <usg>Hdt.</usg>, attic</sense>
<sense level="3" n="2" id="n1401.2">of persons, <trans><tr>truthful</tr></trans>, <usg>Il.</usg>, attic</sense>
<sense level="3" n="3" id="n1401.3">of oracles and the like, <trans><tr>true, coming true</tr></trans>, <usg>Aesch.</usg>, etc.</sense>
<sense level="2" n="II" id="n1401.4">adv. <ref targOrder="U" lang="greek">a)lhqw=s</ref>, ionic <foreign lang="greek">-qe/ws</foreign>,
  <trans><tr>truly</tr></trans>, <usg>Hdt.</usg>, etc.</sense>
<sense level="3" n="2" id="n1401.5"><trans><tr>really, actually, in reality</tr></trans>, <usg>Aesch.</usg>, <usg>Thuc.</usg>, etc.; so,
  <foreign lang="greek">w(s a)lhqw=s</foreign> <usg>Eur.</usg>, <usg>Plat.</usg>, etc.</sense>
<sense level="2" n="III" id="n1401.6">neut. as adv., proparox. <foreign lang="greek">a)/lhqes</foreign>;
  <trans><tr>itane? indeed? really? in sooth?</tr></trans> ironically, <usg>Soph.</usg>, <usg>Eur.</usg>, etc.</sense>
<sense level="3" n="2" id="n1401.7"><foreign lang="greek">to\ a)lhqe/s</foreign> <trans><tr>really and truly</tr></trans>, Lat.
  <foreign lang="la">revera</foreign>, <usg>Plat.</usg>, etc.; so, <foreign lang="greek">to\ a)lhqe/staton</foreign>
  <trans><tr>in very truth</tr></trans>, <usg>Thuc.</usg> </sense>
</entry>
```

Fig 2. The TEI-XML encoding of the ML entry for ἀληθής.

The TEI-XML encoding for each entry already contains most of the information which we wish to represent in RDF marked up, and so the actual processing of the dataset was fairly straightforward. The part of the conversion which, however, did call for some thought was the use of the *lemon* model to structure the RDF translation.

Translating the Middle Liddell in RDF using *lemon*

For the conversion of the TEI-XML version of the ML we decided to use the *lemon* model for publishing lexicons in RDF (McCrae et al 2011, McCrae et al 2012). *lemon* has by now become a de facto standard for converting lexicographic resources into RDF and has been used to convert a number of important lexical resources such as Wordnet (McCrae 2014), UBY (Eckle-Kohler et al 2014), Wiktionary (McCrae et al 2012), and Parole/Simple/Clips (Del Gratta et al 2015). The Linked Open Data movement emphasises the re-use of general vocabularies and models in order to ensure semantic interoperability between datasets. And so given the widespread use of *lemon* in converting lexical resources into RDF, and given the lack of a more specific alternative specifically tailored for lexicographic resources like the ML, we decided to use it as the framework for our conversion. However using *lemon* for the conversion has not been without its challenges. One of the primary difficulties rests in the fact that *lemon* was originally intended as an onomasiological model, that it is, it was designed with the perspective in mind of enriching an already existing ontological or conceptual resource with linguistic information (Cimiano et al 2013). However in our case we started out with a very rich lexical resource but without any particular pre-ordained ontological or conceptual datasets in mind to which to link it. In fact we are not currently using the lemon:reference relation to link our dataset to others.

The ML in particular and the LSJ more generally represent lexical resources that have a specific and relatively complex way of encoding information and that contain a lot of philological and historical data alongside or in addition to "pure" semantic information. Therefore in in order to ensure a faithful translation we had to define a number of new classes and relations in addition to those in *lemon*. In what follows below we will briefly describe (most of) the additional classes and properties which were introduced in order to model the ML and which together make up the lemonLSJ module. Fig. 3 below is a diagram showing classes and properties in the lemonLSJ module and their relation to some of the main classes.

The new class lemonLSJ:Gloss represents the written text associated with each sense; the object relation lemonLSJ:gloss then links elements of this class to lemon senses. The lemonLSJ:usage relation links a sense to an ontological resource describing where that sense was used, e.g., by linking to an author or work where that sense can be found.[1] We have also added the relations lemonLSJ:senseChild, and lemonLSJ:senseSibling in order to represent the nesting of subsenses in ML entries.

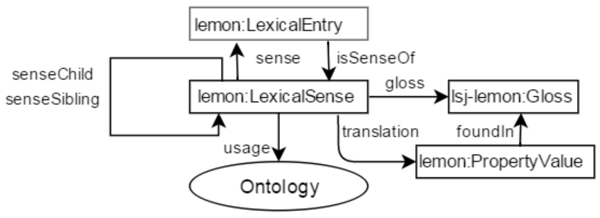
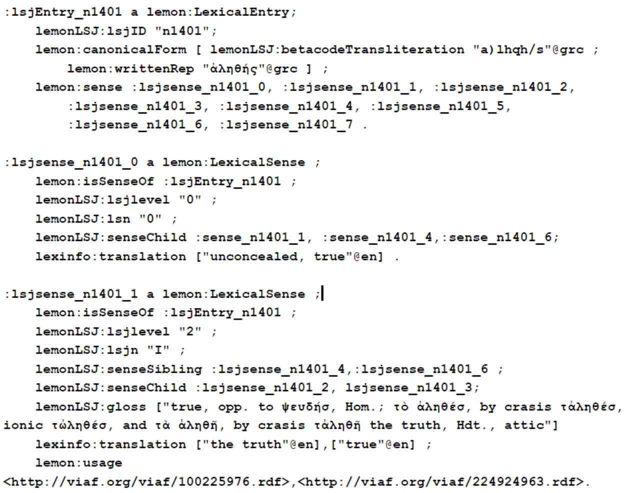


Fig. 3 The lemonLSJ module.

In Fig 4 we present an excerpt of the lemonLSJ RDF-Turtle encoding of the ML entry for ἀληθής with only two of the attached senses represented. Note that we have linked the second of the word senses in Fig 4 to instances in the VIAF dataset (in this case the entries for Herodotus and Homer).



```
:lsjEntry_n1401 a lemon:LexicalEntry;
    lemonLSJ:lsjID "n1401";
    lemon:canonicalForm [ lemonLSJ:betacodeTransliteration "a)lhqh/s"@grc ;
        lemon:writtenRep "ἀληθής"@grc ] ;
    lemon:sense :lsjsense_n1401_0, :lsjsense_n1401_1, :lsjsense_n1401_2,
        :lsjsense_n1401_3, :lsjsense_n1401_4, :lsjsense_n1401_5,
        :lsjsense_n1401_6, :lsjsense_n1401_7 .

:lsjsense_n1401_0 a lemon:LexicalSense ;
    lemon:isSenseOf :lsjEntry_n1401 ;
    lemonLSJ:lsjlevel "0" ;
    lemonLSJ:lsn "0" ;
    lemonLSJ:senseChild :sense_n1401_1, :sense_n1401_4,:sense_n1401_6;
    lexinfo:translation ["unconcealed, true"@en] .

:lsjsense_n1401_1 a lemon:LexicalSense ;
    lemon:isSenseOf :lsjEntry_n1401 ;
    lemonLSJ:lsjlevel "2" ;
    lemonLSJ:lsjn "I" ;
    lemonLSJ:senseSibling :lsjsense_n1401_4,:lsjsense_n1401_6 ;
    lemonLSJ:senseChild :lsjsense_n1401_2, lsjsense_n1401_3;
    lemonLSJ:gloss ["true, opp. to ψευδήσ, Hom.; τὸ ἀληθέσ, by crasis τἀληθέσ,
ionic τὠληθέσ, and τὰ ἀληθῆ, by crasis τἀληθῆ the truth, Hdt., attic"]
    lexinfo:translation ["the truth"@en],["true"@en] ;
    lemon:usage
<http://viaf.org/viaf/100225976.rdf>,<http://viaf.org/viaf/224924963.rdf>.
```

Fig. 4 Excerpt of the RDF encoding of an ML entry.

The diagram in Fig. 5 shows how the nesting of senses is treated in the RDF encoding.

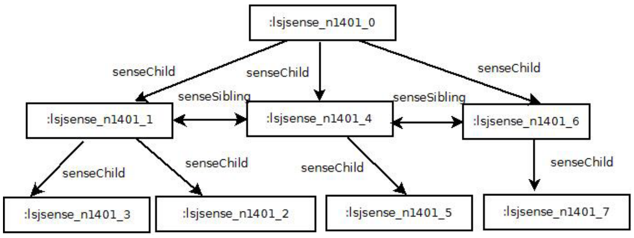


Fig. 5 Sense diagram.

Conclusions and future work

In this paper we have briefly described ongoing work in the conversion of an ancient Greek-English lexicon into RDF. Currently we are manually checking the RDF triples resulting from our conversion scripts and we also plan to additionally link the senses to Princeton Wordnet synsets by checking similarity of the glosses to synset glosses. The resulting dataset will soon be made available as an RDF dump together with other lexical resources of ILC CNR.

Acknowledgments

First of all we would like to thank Perseus project for converting the data and for making it freely available. Thanks also to the organisers of the *mylider* 2015 Summer Datathon, during which some of this work was carried out, for their advice.

Bibliography

Bennett, R., Hengel-Dittrich, C., O'Neill, E. T. and Tillett, B. B. (2006). Viaf (virtual international authority file): Linking die deutsche bibliothek and library of congress name authority files. *World Library and Information Congress: 72nd IFLA General Conference and Council*. Citeseer http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.5328&rep=rep1&type=pdf (accessed 29 October 2015).

Bizzoni, Y., Boschetti, F., Del Gratta, R., Diakoff, H., Monachini, M. and Crane, G. (2014). The Making of Ancient Greek WordNet. *Proceedings of Language Resources and Evaluation Conference, Iceland*. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1071_Paper.pdf (accessed 29 October 2015).

Chiarcos, C., Hellmann, S. and Nordhoff, S. (2011). Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. *TAL*, **52**(3): 245–75.

Cimiano, P., McCrae, J., Buitelaar, P. and Montiel-Ponsoda, E. (2013). On the Role of Senses in the Ontology-Lexicon. In Oltramari, A., Vossen, P., Qin, L. and Hovy, E. (eds), *New Trends of Research in Ontologies and Lexical Resources*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 43–62. http://link.springer.com/10.1007/978-3-642-31782-8_4 (accessed 31 October 2015).

Crane, G., Almas, B., Babeu, A., Cerrato, L., Krohn, A., Baumgart, F., Berti, M., Franzini, G. and Stoyanova, S. (2014). Cataloging for a Billion Word Library of Greek and Latin. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. (DATeCH '14). New York, NY, USA: ACM, pp. 83–88. doi:10.1145/2595188.2595190. http://doi.acm.org/10.1145/2595188.2595190 (accessed 29 October 2015).

Del Gratta, R., Frontini, F., Khan, F. and Monachini, M. (2015). Converting the parole simple clips lexicon into rdf with lemon. *Semantic Web Journal*, **6**(4): 387–92.

Eckle-Kohler, J., McCrae, J. and Chiarcos, C. (2014). LemonUby-a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, Submitted. Special Issue on Multilingual Linked Open Data* http://www.semantic-web-journal.net/system/files/swj404.pdf (accessed 29 October 2015).

Elliott, T. and Gillies, S. (2009). Digital geography and classics. *Digital Humanities Quarterly*, **3**(1).

Francopoulo, G. (2013). *LMF Lexical Markup Framework*. John Wiley & Sons.

Hyvönen, E. (2012). Publishing and Using Cultural Heritage Linked Data on the Semantic Web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, **2**(1): 1–159. doi:10.2200/S00452ED1V01Y201210WBE003.

Janowicz, K. (2009). The Role of Place for the Spatial Referencing of Heritage Data. *The Cultural Heritage of Historic European Cities and Public Participatory GIS Workshop.*. The University of York, UK, pp. 17–18.

Liddell, H. G. and Scott, R. (1896). *An Intermediate Greek-English Lexicon: Founded upon the Seventh Edition of Liddell and Scott's Greek-English Lexicon*. Harper & Brothers.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J. et al. (2012a). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, **46**(4): 701–19.

McCrae, J., Fellbaum, C. and Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. *Proceedings of the 3rd Workshop on Linked Data in Linguistics*. http://pub.uni-bielefeld.de/download/2732779/2732785 (accessed 29 October 2015).

McCrae, J., Montiel-Ponsoda, E. and Cimiano, P. (2012b). Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics*. Springer, pp. 25–34. http://link.springer.com/chapter/10.1007/978-3-642-28249-2_3 (accessed 29 October 2015).

McCrae, J., Spohr, D. and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *The Semantic Web: Research and Applications*. Springer, pp. 245–59. http://link.springer.com/chapter/10.1007/978-3-642-21034-1_17 (accessed 29 October 2015).

Minozzi, S. (2009). *The Latin Wordnet Project*. http://www.dfll.univr.it/documenti/Iniziativa/dall/dall036637.pdf (accessed 29 October 2015).

Oomen, J., Baltussen, L. B. and Erp, M. van (2012). Sharing cultural heritage the linked open data way: Why you should sign up. *Museums and the Web*.

Notes

[1] We made the decision not to include CTS URNs describing particular works or copora, since this data is usually not included in the LM. This data is however available in the Perseus version of the LSJ (though not the ML) and we plan to include it when we come to convert the LSJ into RDF.

Digital Criticism Platform for Evidence-based Digital Humanities with Applications to Historical Studies of Silk Road

Asanobu Kitamoto
kitamoto@nii.ac.jp
National Institute of Informatics, Japan

Yoko Nishimura
nishimura@toyo-bunko.or.jp
Hanazono University, Japan

Introduction

Source criticism is a scholarly process fundamental to many disciplines of humanities, especially in historical studies. However, it is mainly designed for a traditional way of research, namely human scholars read a textual source without producing structured evidences for reuse. Our proposal is to extend the traditional methodology of source criticism to digital research infrastructure on which scholars records the reasoning process from evidences to facts, and share them with other scholars so that the detail of the reasoning process can be transparently reproduced. Our approach, digital criticism, aims at realizing this idea on digital criticism platform (DCP) toward evidence-based digital humanities with the support of Semantic Web technology.

An evidence-based approach is also used in quantitative humanities, but digital criticism is a fundamentally different approach. Digital criticism focuses on reading sources in a critical manner, while quantitative humanities focuses on deriving numerical values from the collection of sources. Generally speaking, quantitative humanities tries to make abstraction of corpus through quantitative aggregation, while digital criticism tries to make critique of sources through digital analysis with support for the management of a layered abstraction process.

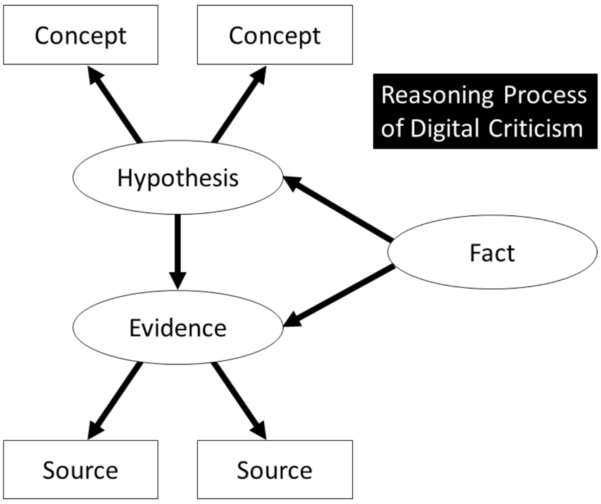


Figure 1 Schematic diagram of evidence network

Core concepts of digital criticism

The main contribution of the paper is evidence network on which digital criticism is performed. The network consists of four concepts, namely evidence, hypothesis, fact, and reliability. Those concepts have our own definition to organize the historical reasoning process into an explicit model.

1. Evidence is relationship between sources. If a photograph A takes the same scene with a photograph B, they are linked as an evidence with reproducible parameters of how photographs can be matched.

2. Hypothesis is relationship between historical concepts. If a ruin A and ruin B have different names but are believed to be the same ruins, they are linked as a hypothesis with supporting evidences and other descriptions on the reasoning process.

3. Fact is relationship between evidences and hypotheses to claim reusable knowledge for future research.

4. Reliability is an attribute of evidences and hypotheses to represent the degree of reliability estimated by the scholar on registration. Because the estimated reliability may be different for each scholar, evidences and hypotheses should always be linked to a user entity who made the action.

Figure 1 illustrates a schematic diagram of evidence network. A scholar can build up the network starting from each evidence and hypothesis. Another scholar who wants to reuse the knowledge can start from the fact, and track back to hypotheses and evidences to check the reliability of the reasoning process.

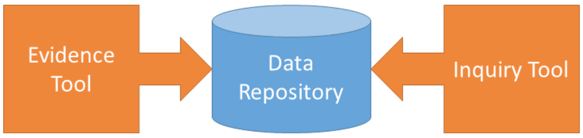


Figure 2 Three components of Digital Criticism Platform (DCP)

Digital Criticism Platform (DCP)

Evidence network is a directed graph with semantic annotation, and the graph structure is built using RDF (Resource Description Framework). Hence a search over a graph can be implemented using SPARQL, which is a query language for a RDF graph. To construct and study the evidence network, we develop Digital Criticism Platform (DCP) with three components, namely data repository, evidence tool, and inquiry tool as illustrated in Figure 2.

1. *Data repository* archives digital resources with unique IDs and metadata. We developed data repository using DSpace as infrastructure for its reliability and extensibility to Semantic Web environment.

2. *Evidence tool* works on various types of media for collecting evidences. We developed three tools, Mapping, Photofit, and MemoryHunt, for maps, photographs, and field work, respectively.

3. *Inquiry tool* is to study and sophisticate the evidence network using Semantic Web technology such as SPARQL. This component is now under study, and has not reached the development phase.

In the following, we focus on three evidence tools to describe more detail of the tools. Those tools should be integrated into data repository so that every evidence is registered as a part of evidence network. Three tools are already in operation, but only Photofit is fully integrated into data repository at the time of writing.

Mappining

Mappining (Kitamoto 2012) is a web-based tool for matching maps (Figure 3). This tool employs an idea of "interactive georeferencing" in contrast to geometric correction. Interactive georeferencing uses a pin to match two maps at a single point, in a similar way of pinning cloths. When we put a pin, two maps scroll together, while when we release a pin, only one map scrolls. Using this interface, any point on a map can be matched with another map for overlay-based comparison. Interactive georeferencing is advantageous for reading sources because geometric distortion is not introduced. A pin is an evidence to claim that each point on a map represent the same location on earth.

Photofit

Photofit is a web-based tool for matching photographs (Figure 4). The target of the tool is two photographs taken from similar locations but different angles at different time. It allows planar shift and zooming transformation for two photographs to find the best match. On success, it is an evidence to claim that two photographs take the same scene or the same object. This evidence may lead to a new hypothesis on the identity of historical concepts when the captions of two photographs represent different historical concepts.

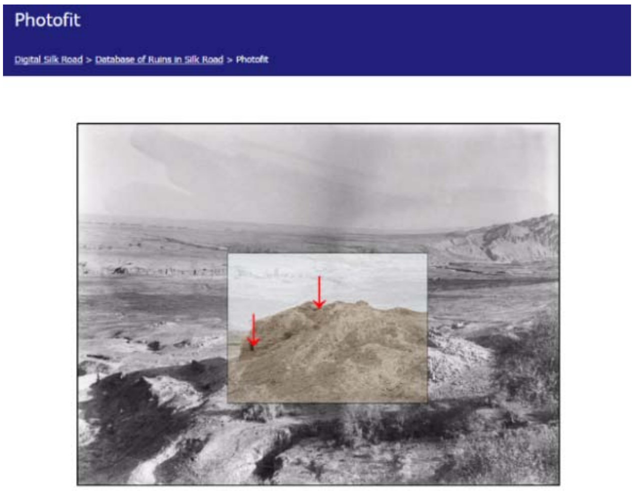

Photofit
Digital Silk Road > Database of Ruins in Silk Road > Photofit

Figure 4 Matching two photographs on Photofits

MemoryHunt

MemoryHunt (Kitamoto 2015) is a mobile app for matching a photograph with the real world (Figure 5). It shows a target photograph on the viewfinder of a smartphone camera with controllable transparency, and the task of a user is to find the same location and the same direction that the photograph was taken. On success, a mobile app can record the location and direction as metadata of the photograph. This may lead to an evidence between an object appearing in the photograph and one in the present world.

Evidence network

The purpose of digital criticism is to build an evidence network in which historical sources are linked through evidence nodes, and historical concepts are linked through hypothesis nodes to derive fact nodes supported by evidence and hypothesis nodes. We start by a bottom-up process of registering evidences and hypotheses which are still fragmented. We then switch to a top-down process of viewing a graph structure as a whole to discover unknown relationships. In the following, we use our past results on Silk Road ruins as case studies to check the validity of our approach. Due to the incompleteness of DCP, the following diagrams were manually drawn.

Figure 5 shows a simple evidence network. A photograph in a book is matched with another photograph in another book using Photofit, and an evidence node is added with transformation parameters. We also know that, through captions, each photograph represents a ruin known by different names. We then add a hypothesis node based on the above evidence to claim that two ruins known by different names are the same.

Figure 6 shows a complex evidence network. Multiple tools are used for matching multiple sources, such as Mappinning, Photofit, and Google Earth. The evidence network suggests relationship among a Buddhist ruin

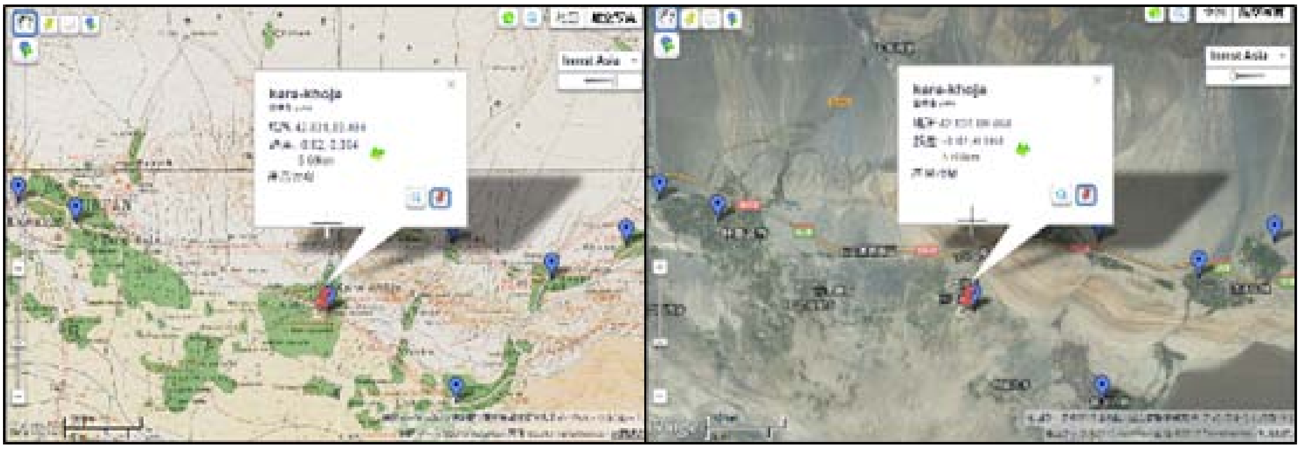


Figure 3 Matching two maps on Mappining

and a ruin known by the name Chikkan-kol and another ruin known by "七康湖遺跡." This relationship was our discovery previously reported in a paper (Nishimura and Kitamoto 2010), but digital publishing in the form of evidence network offers clearer representation of a compex reasoning process.

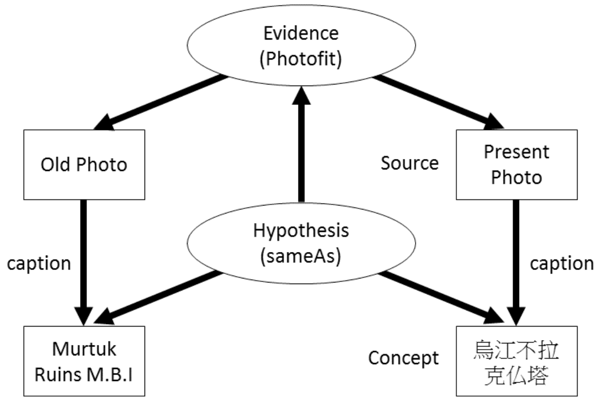


Figure 6 A simple evidence network

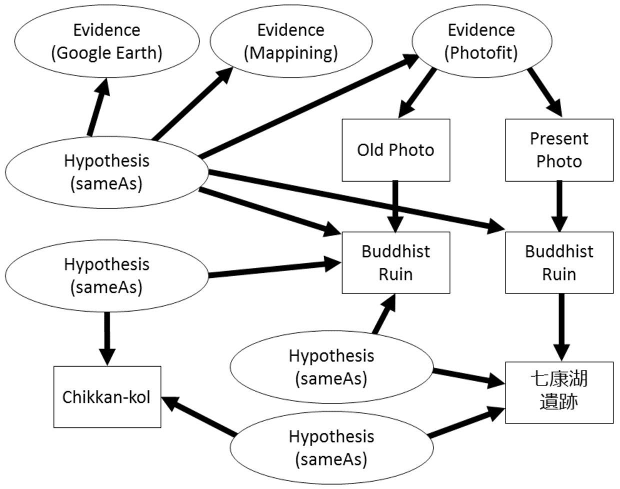


Figure 7 A complex evidence network

Conclusion

We proposed digital criticism platform (DCP) as a model of source criticism on a digital platform. Digital criticism tries to simulate the traditional method of source criticism and extend it to take advantage of digital research infrastructure. Digital Criticism is the upgraded version of our past proposal on "data criticism" (Kitamoto and Nishimura 2014), after shifting research focus on the type of sources to the way of digital scholarship. The methodology of criticism is also investigated in different approaches, such as algorithmic criticism (Ramsay 2011).

Knowledge representation of historical evidences has a large number of related literature. In particular, Pasin proposed the usage of factoid model for accumulating evidences for prosopography in the linked data world (Pasin and Bradley 2013), and linked data is also used for places such as Pelagios (Isaken et al. 2014). We also use the same concept of linking entities and uses sources as evidences, but digital criticism focuses more on accumulating structured evidences and hypothetical links. Another important work for the refinement of evidence network is knowledge representation for argumentation, such as CRMinf argumentation model, an extension of CIDOC-CRM (Paveprime Ltd and collaborators 2015). Finally, we are yet to make choices on metadata standards or vocabularies, which are open questions.

The long-term goal is to design a digital publishing platform for evidence citation. Historical facts can be tracked to evidences to clarify the reliability of evidences that support historical facts. This leads to increased transparency of research, and to realize better data management planning.

Acknowledgments

This work is supported by JSPS KAKENHI Grant Number 26540178. Software codes for DCP was developed by Mr. Tomohiro Ikezaki.

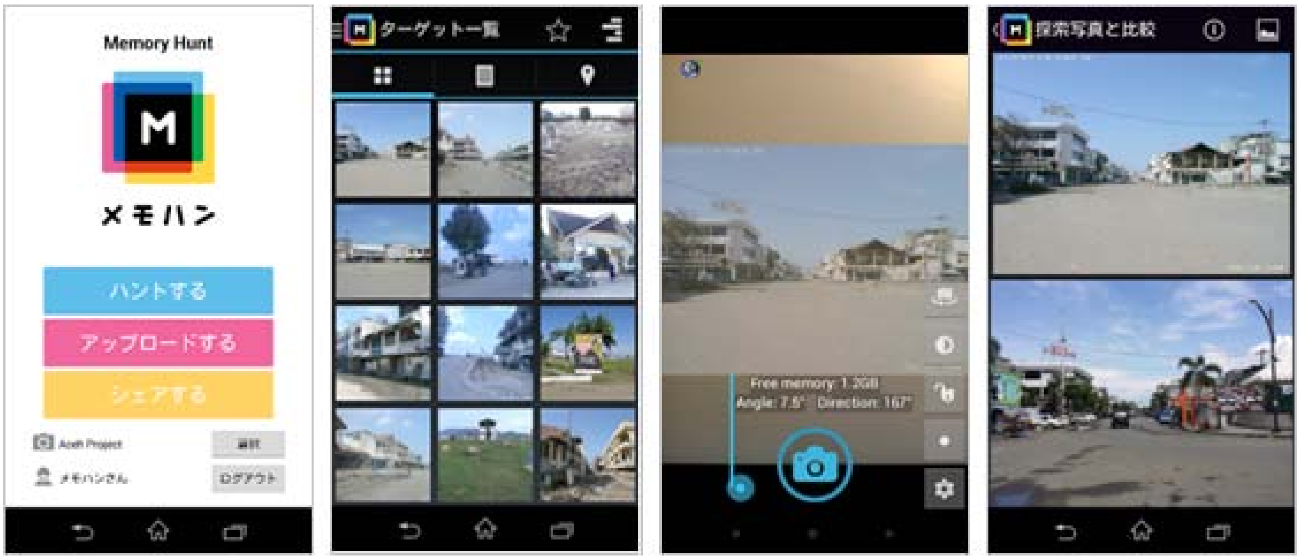


Figure 5 Interface of MeomoryHunt app

Bibliography

Isaken, L., Simon, R., Barker, E. T. E. and Cañamares, P. S. (2014). Pelagios and the emerging graph of ancient world data. *Proceedings of the 2014 ACM conference on Web science*, pp. 197-201.

Kitamoto, A. (2012). *Mappining.* http://dsr.nii.ac.jp/digital-maps/mappinning/ (accessed 6 March 2016).

Kitamoto, A. and Nishimura, Y. (2014). Data Criticism: General Framework for the Quantitative Interpretation of Non-Textual Sources. *Digital Humanities 2014: Conference Abstracts.*

Kitamoto, A. (2015). MemoryHunt: A Mobile App with an Active Viewfinder for Crowdsourced Annotation through the Re-experience of the Photographer. *Fifth Annual Conference of the Japanese Association for Digital Humanities (JADH2015).*

Nishimura, Y. and Kitamoto, A. (2010). Identification of Ruins Excavated by Silk Road Expeditions through Matching Names and Locations by Stein Maps and Google Earth. *IPSJ SIG Computers and the Humanities Symposium 2010*, pp. 255-62. (in Japanese).

Pasin, M. and Bradley, J. (2013). Factoid-based prosopography and computer ontologies: towards an integrated approach. *Digital Scholarship in the Humanities*, **30**(1): 86-97.

Paveprime Ltd and collaborators. (2015). CRMinf: the Argumentation Model, An Extension of CIDOC-CRM to support argumentation.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press.

Museum Digitization Practices Across Russia: Survey and Web Site Exploration Results

Inna Kizhner
inna.kizhner@gmail.com
Siberian Federal University, Russian Federation

Melissa Terras
m.terras@ucl.ac.uk
University College London, UK

Maxim Rumyantsev
m-rumyantsev@yandex.ru
Siberian Federal University, Russian Federation

Introduction

An appeal to strengthen DH research using audiovisual collections has been repeatedly articulated in the DH community. Some of these collections are digitized museum collections of heritage objects. Many of these serve as repositories (sometimes for internal inventory purposes), some of the collections follow the rules of building digital scholarly editions and provide tools of analysis/summarizing and, sometimes, sources for critical interpretation.

Little is known, however about digitization practices within Museums in Russia, with no prior research into the number of Russian museum web sites, the amount of their collections digitized or put online or their digitization procedures. International research on digitization in museums has a comparatively long tradition and there is a significant literature on statistical data about digitization, digital preservation and online access to cultural heritage across Europe (ENUMERATE, 2015) and Northern America. Reports on digitization success stories in these geographical areas have been published (see among others ENUMERATE, 2015; Clough, 2013; Olsen, 2015) but there is no prior research regarding digitization uptake and practices in Russia.

This paper explores Russian museum digitization practices employing the standard method of surveying museums employed by the ENUMERATE project, allowing us to compare results to ENUMERATE. We also chose to augment our findings with the results from exploring the content of museum web sites to understand what parts of museum analogue collections are posted online. The paper seeks to answer a critical question about the size of digital collections as measured by the ratio of digital copies of unique museum objects to the number of unique objects from a museum collection (or the number of digital copies posted on the museum web site). We understand it very well that this can only serve as an imperfect proxy for digitization practices in Russian museums. However, in a situation when this is the only data that could be obtained, we judged it would be reasonable to start the discussion of digitization in Russian museums from this point.

Methods

In the early stages of our project we adapted the ENUMERATE questionnaire to our goals of obtaining answers from museum staff. We posted a survey with twelve questions online in October 2014 and we sent letters to 440 museums with a link to the survey. The email addresses were obtained from Museums of Russia web portal.[1] The database lists 3063 Russian museums including data on the number of visitors per year, the year when a museum was established and the number of curators among its employees. Our sampling method was to choose 130 museums located in Moscow and Saint Petersburg and 310 museums located in provincial cities and smaller settlements. Each of the 80 administrative districts in the Russian Federation was represented by 3-6 museums with one or two of them belonging to the group with the number of visitors per year more than 50,000 people. The other two groups included small museums (the number of visitors per year was fewer than 15,000 people) and medium museums with the number of visitors between 15,000 and 50,000 per year. This gave us an appropriate sample of museums to begin to understand different museum digitization practices across Russia.

Our next step was to study the web sites of large provincial museums and medium-sized provincial museums for 58 administrative districts (116 provincial museums as a total). Nineteen web sites for large museums and 23 web sites for medium-sized museums in Moscow and Saint Petersburg were also studied. The number of digital images on museum web sites was compared with the number of unique objects in their collections as reported in the Museums of Russia database (Museums of Russia, 2015).

Results and discussion

The response rate for the survey was a disappointing 6% (30 memory institutions completed the questionnaire and answered the most important question about the size of their digital collection). Such a response rate was very low compared to 30% response rate for the survey of library digitization projects in the USA in 2004 (Boock and Vondracek, 2006) and 51% for European cultural institutions in May 2007 and May 2009 (Poll, 2010). This result though is consistent with the finding that other Russian surveys tend to demonstrate low response rates with some studies reporting low level of trust to surveys among respondents (Kalinin, 2012), which has important methodological implications for those carrying out research within Russia.

In Table 1 we summarize the results for the survey of museum digital collections. Absolute average of the proportion of an analogue collection that was reported digitized for the 30 museums in the sample was 18,3%, in line with the results from ENUMERATE survey for 2012 (ENUMERATE, 2014).

ratio of digital images to unique objects (%)	number of museums in a sample	as % of a number of museums in a sample
0	8	26.3
10	8	26.3
20	8	26.3
30	3	10
50	1	3.3
80	1	3.3
90	1	3.3

Table 1. Distribution of parts of collection digitized (as reported by museums in our survey)

Tables 2, 3, 4, and 5 show the results for exploring the web sites of 158 Russian museums in the provinces and major cities to find out what parts of their analogue collections are posted online.

% of digital images representing museum collections online	number of museums	as % of large provincial museums in the sample
0	8	14
from 0 to 0,1	27	47
from 0,1 to 0,98	14	25
from 1 to 10	6	10
more than 10	2	3

Table 2. Parts of museum analogue collections published online. Large provincial museums.

% of digital images representing museum collections online	number of museums	as % of medium-sized provincial museums in the sample
0	13	22
from 0 to 0,1	11	19
from 0,1 to 0,98	21	36
from 1 to 10	9	15
more than 10	4	7

Table 3 Parts of museum analogue collections published online. Medium-sized provincial museums.

% of digital images representing museum collections online	number of museums	as % of museums in the sample
0	6	14
from 0 to 0,1	10	23
from 0,1 to 0,98	15	35
from 1 to 10	11	26
more than 10	1	2

Table 4 Parts of museum analogue collections published online. Museums in Moscow and Saint Petersburg

% of digital images representing museum collections online	number of museums	as % of museums in the sample
0	27	17
from 0 to 0,1	48	30
from 0,1 to 0,98	50	31
from 1 to 10	26	16
more than 10	7	4

Table 5 Parts of museum analogue collections published online. Overall results for 158 museums in the sample.

As shown in Table 5, a third of Russian museums in our sample publish less than 0,1% of their images online while another third of museums post digital images for a bigger part of their collection (but still less than 1%).

Large provincial museums are not enthusiastic about publishing their images online, with half of studied web sites demonstrating results which were lower than 0,1% of their analogue collections. Medium-sized museums show slightly better results, with a third of them displaying between 0,1 to 1% of their analogue collections online. They, however, have fewer objects to digitize and annotate. Museums in the two major cities seem more inclined to post their images online, with a quarter of museums publishing between 1% and 10% of their analogue collections on the Web.

Limitations

Unfortunately, the survey was designed so that respondents only had an opportunity to choose between 0% and 10% options when describing the parts of their analogue collections being digitized. This may have left many respondents with collections in between these figures indecisive on what option to choose, deteriorated the quality and accuracy of the results and may have influenced the response rate. The low response rate is also problematic, but, when combined with our survey of practice still gives us interesting insights to a hitherto undocumented area. Future work will be needed to work with museum bodies on gathering further data.

Conclusion

This work includes the survey results of thirty Russian museum digital collections to find out what part of their analogue collections is digitized. We also studied 158 museum web sites to count the number of digital images representing museum objects and to compare this number to the number of objects in the analogue collections. The average ratio of Russian museum digital collections in the sample of 30 museums compared to their analogue collections was 18,3% which is in line with the results from the ENUMERATE project (ENUMERATE, 2014), and aligns the work of Russian museums to the rate of digitization across Europe.

We have shown that half of large provincial museums in Russia publish an insignificant number of their images on the Web, and further work will allow us to establish why this is the case. Also (at a time when European Museums are being encouraged to adopt the Open Licensing Agenda), further work will pursue the opportunities of sharing digitized collections within Russian legislation.

Bibliography

Boock, M. and R. Vondracek. (2006). *Organizing for Digitization: A Survey, portal: Libraries and the Academy*, **6**(2): 197-217.

Clough, G. W. (2013). *Best of Both Worlds: Museums, Libraries, and Archives in a Digital Age.* Washington, DC: Smithsonian Institution. http://www.si.edu/content/gwc/BestofBothWorldsSmithsonian.pdf (accessed 20 October 2015).

ENUMERATE. (2015). *ICT Policy Support Programme of the European Commission.*http://www.enumerate.eu (accessed 20 October 2015).

ENUMERATE STATISTICS. (2014). *ICT Policy Support Programme of the European Commission.* http://www.enumerate.eu/en/statistics/ (accessed 20 October 2015).

Kalinin, K. (2012). Open opinion: a new hope, *Sociological Journal*, **1**: 167-71. In Russian. http://jour.isras.ru/index.php/socjour/article/viewFile/459/433 (accessed 20 October 2015).

Mikhailovskaya, A. and K. Nasedkin. (2002). *The Museums of Russia Web Portal. Museum International*, **54**(4): 52-56.

Museums of Russia. (2015). *Russian Network of Cultural Heritage.* http://www.museum.ru (accessed 20 October 2015).

Olsen, E. (2015). Museum Specimens Find New Life Online. *The New York Times*, October 19. http://www.nytimes.com/2015/10/20/science/putting-museums-samples-of-life-on-the-internet.html?emc=eta1&_r=0 (accessed 20 October 2015).

Poll, R. (2010). NUMERIC: statistics for the digitization of European cultural heritage, *Electronic Library and Information Systems*, **44**(2): 122-31.

Notes

[1] Museums of Russia web portal (http://www.museum.ru) includes detailed information on 3063 Russian museums, ranks their web sites, posts news, discussion threads and announcements for curators. Online since 1996, it was initiated by the State Darwin Museum and supported by the RF Ministry of Culture (see also Mikhailovskaya and Nasedkin 2002).

Tracking Online User Behaviour With A Multimethod Research Design

Martijn Kleppe
m.kleppe@vu.nl
Vrije Universiteit Amsterdam, Netherlands, The

Otte Marco
m.otte@vu.nl
Vrije Universiteit Amsterdam, Netherlands, The

Understanding users' online behaviour is of growing interest to academic researchers in a variety of fields. Traditionally, in the marketing domain commercial research companies map consumer behaviour to understand when and where customers decide to buy products. For this purpose, web metrics of individual websites serve as detailed source of information on when, how and at which section a user enters a website. Recently this type of data is also being used by cultural heritage institutes to understand the interest of their visitors (De Haan and

Adolfsen, 2008), to track where their digital content is being reused (Navarrete Hernández, 2014) or to understand the query's users perform in search systems by analysing the log files (Batista and Silva, 2002; Huurnink, 2010). In this type of research, the website is the central research object providing traces that Menchen-Trevino (2013) calls 'Horizontal Data sets'. These contain data that are 'organized around a specific type of trace, for example search terms, web browsing log files, tweets, hashtags, likes or friend and follower ties' (Menchen-Trevino, 2013: 331). An advantage of using this type of data is that they are not obtrusive to the respondents since they are created automatically as users are surfing the web. However, this also leads to an ethical disadvantage since users are not aware that their online behaviour is being examined, nor could they give their consent to have their data being analysed. While Horizontal Data Sets are organized around one type of trace, Vertical Data Sets are organized around research participants that deliberately 'give permission for researchers to collect their digital traces' (Menchen-Trevino, 2013: 331).

Since mid-1990s, commercial research agencies have started to collect these types of vertical data by building tools and panels of respondents whose online behaviour is monitored 24/7 to provide data on usage across media and purchase behaviour (Coffey, 2001; Napoli, 2010; Taneja and Mamoria, 2012). Similar to television viewing rates, these lists are mainly created to gain more insight in the background of website visitors in order to provide potential advertisers with information on how to reach their online target audience in the best possible manner. Obviously these commercial research data contain very rich information, also for academics who are interested in collecting real-world Web use data. However, apart from lists of the most popular domains that are published as open data by companies such as Alexa and Similarweb[1], data containing information about visits to each individual page and information about the background of the panel is not available. Main arguments of commercial agencies to not collaborate with scholars is to ensure the confidentiality of their respondents' identity and to prevent scholars to gain insight into the techniques applied by the companies.

Nevertheless, researchers in a variety of disciplines are interested in tracking online behaviour in a real-world situation. Especially in the communications science realm, scholars experimented with several techniques of tracking people's online behaviour (Ebersole, 2000; Tewksbury, 2003; Findahl et al., 2013; Findahl, 2009; Munson et al., 2013; Damme et al., 2015; Menchen-Trevino and Karr, 2012). Striking about these pioneering monitoring studies is its multi-method approach. By default each does not only monitor web use but also compares its outcome with either survey, diary or interview data. By triangulating the results, these researchers try to overcome the critique on classic studies on media consumption that often deploy either surveys or diaries registering self-reported media behaviour (e.g. Van Cauwenberge, d'Haenens and Beentjes, 2010; Schrøder and Kobbernagel, 2010; Taneja et al., 2012; Reuters Institute for the Study of Journalism, 2015). These methods strongly rely on the memory of the participants while several scholars found respondents often overestimate their media use (Ebersole, 2000; Prior, 2009; Robinson, 1985). Furthermore, since filling in diaries and surveys on news consumption is very labour-intensive, its outcomes mainly focus on *when* media have been consumed or on *which* devices, while it remains unclear *what* has been consumed. One way to gain insight in the consumed news content is focus on metrics of individual news organisations (Batista and Silva, 2002; Boczkowski et al., 2011; Lee et al., 2012; Usher, 2013) or the most clicked items (e.g. Boczkowski et al., 2011; Karlsson and Clerwall, 2013; Lee et al., 2012; Nederlandse Nieuwsmonitor, 2013). However, given the focus of these studies on individual websites or most-clicked articles it remains unknown which genres of news websites constitute users' 24/7 news menu. Do they e.g. visit news about sports mainly in the morning and news on politics mainly during the evening? Taneja at el. (2012) tried to overcome this problem by literally following 495 users throughout an entire day. However, even with this labour-intensive fieldwork it proved not to be possible to incorporate the genres of consumed news items.

Web monitoring tools such as the above mentioned examples, now offer a less labour-intensive and more precise way of registering digitally consumed news items. By deploying these techniques, we could overcome the knowledge gap of the 24/7 news consumption menu. Therefore, we created *the Newstracker*, a custom built system that collects web activities of specified and authenticated users, cleans the data by removing non-relevant data, extracts the associated content and stores this as a new dataset to be used for analysis. While most existing online tracking studies mainly report the visited websites, our set-up goes two steps further. We did not only monitor the website titles but also the actual visited URLs and crawled all textual and visual contents of the visited websites. Since one of the problems when monitoring a person's online behaviour is the magnitude of the data that is being collected (Batista and Silva, 2002; Manovich, 2012; Vicente-Marino, 2013: 43), we deployed automated content analyses techniques (Atteveldt, 2008; et al., 2012) to detect the topics that are being discussed in the news items. This enabled us to calculate the topical online news consumption during the day.

In this paper we will describe the set-up of 'The Newstracker' in a study on the online news consumption of a group of young Dutch news users and its applicability for other types of Digital Humanities research such as user studies focussing on formulating requirements based on existing user behaviour. We will demonstrate the workflow of the Newstracker and how we designed the data collection and pre-processing phase (see figure 1).

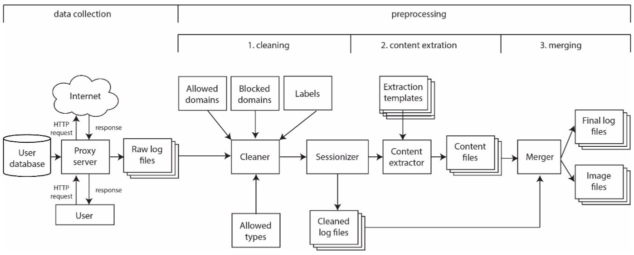


Figure 1. Workflow of the Newstracker application, illustrating the two main phases: Data Collection and Pre-processing. The latter consists of three stages: cleaning, content extraction and merging

By reflecting on the technical, methodological and analytical challenges we encountered, we will illustrate the potential of online monitoring tools such as the Newstracker. We will end our paper with discussing its limitations by stressing the need for a multimethod study design when aiming not only to register but also to understand online user behaviour.

Bibliography

Batista, P. and Silva, M. (2002). Mining Web Access Logs of an On-line Newspape. Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems http://xldb.di.fc.ul.pt/xldb/publications/rpec02.pdf.

Bhulai, S., Kampstra, P., Kooiman, L., Koole, G. and Kok, B. (2012). Trend visualization on Twitter: What's hot and what's not?. IARIA. (Data Analytics). Barcelona, pp. 43–48.

Boczkowski, P. J., Mitchelstein, E. and Walter, M. (2011). Convergence Across Divergence: Understanding the Gap in the Online News Choices of Journalists and Consumers in Western Europe and Latin America. Communication Research, **38**(3): 376–96 doi:10.1177/0093650210384989.

Coffey, S. (2001). Internet Audience Measurement: A Practicioner's View. Journal of Interactive Advertising, **1**(2): 10–17.

Damme, K. V., Courtois, C., Verbrugge, K. and Marez, L. D. (2015). What's APPening to news? A mixed-method audience-centred study on mobile news consumption. Mobile Media & Communication, **3**(2): 196–213 doi:10.1177/2050157914557691.

De Haan, J. and Adolfsen, A. (2008). De Virtuele Cultuurbezoeker. Den Haag: Sociaal en Cultureel Planbureau http://www.scp.nl/dsresource?objectid=19697&.

Ebersole, S. (2000). Uses and Gratifications of the Web among Students. Journal of Computer-Mediated Communication, **6**(1): 0–0 doi:10.1111/j.1083-6101.2000.tb00111.x.

Findahl, O. (2009). The Swedes and the Internet 2009. Gävle: World Internet Institute.

Findahl, O., Lagerstedt, C. and Aurelius, A. (2013). Triangulation as a way to validate and deepen the knowledge about user behavior. A comparison between questionnaires, diaries and traffic measurement. Audience Research Methodologies: Between Innovation and Consolidation. New York, pp. 54–69.

Huurnink, B. (2010). Search in Audiovisual Boradcast Archives Amsterdam: University of Amsterdam http://dare.uva.nl/document/2/83234.

Karlsson, M. and Clerwall, C. (2013). Negotiating Professional News Judgment and "Clicks". Nordicom Review, **34**(2): 65–76 doi:10.2478/nor-2013-0054.

Kleinneijenhuis, J. and Atteveldt, W. van (2006). Geautomatiseerde inhoudsanalyse, met de berichtgeving over het EU-referendum als voorbeeld. Inhoudsanalyse: Theorie En Praktijk. Kluwer, pp. 227–50.

Lee, A. M., Lewis, S. C. and Powers, M. (2012). Audience Clicks and News Placement: A Study of Time-Lagged Influence in Online Journalism. Communication Research: 0093650212467031 doi:10.1177/0093650212467031.

Manovich, L. (2012). How to Follow Software Users? http://lab.softwarestudies.com/2012/04/new-article-lev-manovich-how-to-follow.html (accessed 28 January 2014).

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research. Policy & Internet, **5**(3): 328–39 doi:10.1002/1944-2866.POI336.

Menchen-Trevino, E. Tracing our every move: Big data and multi-method research The Policy and Internet Blog http://blogs.oii.ox.ac.uk/policy/tracing-our-every-move-big-data-and-multi-method-research/ (accessed 30 April 2015).

Menchen-Trevino, E. and Karr, C. (2012). Researching Real-World Web Use with Roxy: Collecting Observational Web Data with Informed Consent. Journal of Information Technology & Politics, **9**(3): 254–68 doi:10.1080/19331681.2012.664966.

Munson, S. A., Lee, S. Y. and Resnick, P. (2013). Encouraging Reading of Diverse Political Viewpoints with a Browser Widget. Seventh International AAAI Conference on Weblogs and Social Media http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6119 (accessed 25 September 2015).

Napoli, P. M. (2010). Audience Evolution: New Technologies and the Transformation of Media Audiences. New York: Columbia University Press http://cup.columbia.edu/book/audience-evolution/9780231150347.

Navarrete Hernández, T. (2014). A history of digitization: Dutch museums Amsterdam: University of Amsterdam http://dare.uva.nl/record/1/433221 (accessed 25 September 2015).

Nederlandse Nieuwsmonitor (2013). Seksmoord Op Horrorvakantie: De Invloed van Bezoekersgedrag Op Krantenwebsites Op de Nieuwsselectie van Dagbladen En Hun Websites. Amsterdam: Nederlandse Nieuwsmonitor http://www.nieuwsmonitor.net/d/244/Seksmoord_op_Horrorvakantie_pdf.

Prior, M. (2009). The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure. Public Opinion Quarterly, **73**(1): 130–43 doi:10.1093/poq/nfp002.

Reuters Institute for the Study of Journalism (2015). Digital News Report 2015. Oxford http://www.digitalnewsreport.org/.

Robinson, J. P. (1985). The validity and reliability of diaries versus alternative time use measures. Time, Goods, and Well-Being.

Schrøder, K. C. and Kobbernagel, C. (2010). Towards a typology of cross-media news consumption: a qualitative-quantitative synthesis. Northern Lights: Film and Media Studies Yearbook, **8**(1): 115–37 doi:10.1386/nl.8.115_1.

Taneja, H. and Mamoria, U. (2012). Measuring Media Use Across Platforms: Evolving Audience Information Systems. International Journal on Media Management, **14**(2): 121–40 doi:10.1080/14241277.2011.648468.

Taneja, H., Webster, J. G., Malthouse, E. C. and Ksiazek, T. B. (2012). Media consumption across platforms: Identifying user-defined repertoires. New Media & Society, **14**(6): 951–68 doi:10.1177/1461444811436146.

Tewksbury, D. (2003). What Do Americans Really Want to Know? Tracking the Behavior of News Readers on the Internet. Journal of Communication, **53**(4): 694–710 doi:10.1111/j.1460-2466.2003.tb02918.x.

Usher, N. (2013). Al Jazeera English Online. Understanding Web metrics and news production when a quantified audience is not a commodified audience. Digital Journalism, **1**(3): 335–51 doi:10.1080/21670811.2013.801690.

Van Cauwenberge, A., Haenens, L. S. J. d' and Beentjes, J. W. J. (2010). Emerging consumption patterns among young people of traditional and internet news platforms in the Low Countries. Observatorio, **4**(3): 335–52.

Vicente-Marino, M. (2013). Audience research methods. Facing the challenges of transforming audiences. Audience Research Methodologies: Between Innovation and Consolidation. New York, pp. 37–53.

Notes

[1] http://www.alexa.com/topsites , http://www.similarweb.com/global

Researchers to your Driving Seats: Building a Graphical User Interface for Multilingual Topic-Modelling in R with Shiny

Thomas Koentges
thomas.koentges@uni-leipzig.de
University of Leipzig, Germany

In this paper I will showcase the results of topic-modelling research undertaken during my 2015 visiting fellowship at Victoria University Wellington (VUW), New Zealand. This research was undertaken in collaboration with staff at the Alexander Turnbull Library, National Library of New Zealand (ATL), and was subsequently applied to the Open Philology Project (OPP) in the Department of Digital Humanities at the University of Leipzig, Germany, and to Classical Persian corpora in collaboration with the Roshan Institute for Persian Studies at the University of Maryland (UMD), USA. The paper emphasizes how humanities researchers, even those who do not have deep scripting knowledge, can be enabled not only to understand, but also to modify the topic-modelling process, adapt it to their corpus- or language-specific needs, and then use the results of this process for further qualitative research. Topic-modelling in digital humanities research is a means to an end that should always help to answer a specific research question. In this way the paper follows an alternative route to previous interactive topic-modelling research, for example Hu et al. (2014) at the University of Maryland. Instead of asking how the method of topic-modelling can be improved by user-input, the author of this paper focuses on the ways in which we can improve the results of topic-modelling while also improving the humanities researcher's understanding of the method and its variables.

After a brief introduction to the research projects in Wellington and Leipzig and to topic-modelling itself, the paper will summarize the limitations of topic-modelling with special emphasis on how to determine an ideal number of topics, as well as a short discussion of morphosyntactic normalization and the use of stop-words. It will then suggest a researcher-focused method of addressing these limitations and challenges in topic-modelling by introducing the Shiny topic-modelling application developed by the author based on R, Shiny, and J. Chang's LDA library and C. Sievert's LDAvis library. The paper will then briefly demonstrate the applicability to the different use-cases at ATL and OPP, which deal with very different fields and languages, including English, Latin, Ancient Greek, Classical Arabic and Persian. The paper will finish by stressing how digital humanities research results and practices can be improved by enabling humanities researchers who focus on more traditional and qualitative analyses of the corpora to use the quantitative method of topic-modelling as a macro-scope and faceting tool; effectively calling humanities researchers back to their driving seats.

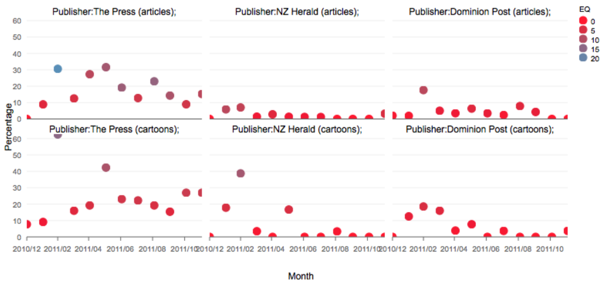


Figure 1: Percentage of earthquake-related articles and cartoons published in three New Zealand newspapers in the period from December 2010 to December 2011 generated using LDA-topic modelling. Also showing combined earthquake score.

During my research stay at VUW I worked with the Research Librarian for cartoons at ATL, Dr Melinda Johnston, on a mixed-methods-based analysis of the reactions of cartoonists and New Zealand print publications to the Canterbury Earthquakes in 2010 and 2011. ATL is part of the National Library of New Zealand, an institution that is interested in making the country's cultural heritage more accessible to a digital audience and researchers. Within the short project I attempted to automatize the detection and analysis of cartoon descriptions created by ATL and over 100,000 abstracts produced by Index New Zealand

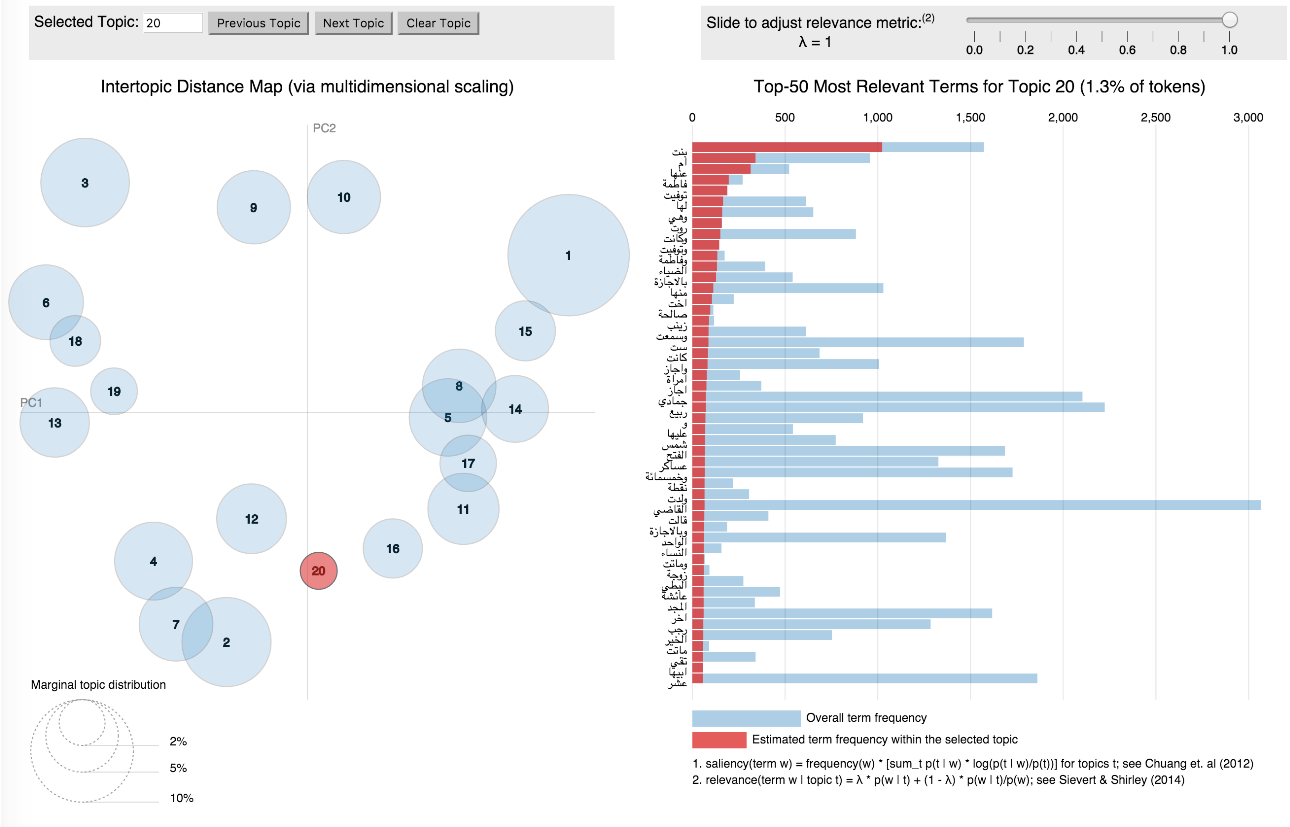
(INNZ); all items were published between September 2010 and January 2014. The INNZ-data could be retrieved as a double-zipped XML file from INNZ's webpage and ATL's item descriptions could be queried using the Digital New Zealand (DNZ) API. During the project it became apparent how a topic-modelling approach, first applied by the author in VUW's Digital Colenso project, could considerably speed up the finding of earthquake-related descriptions and abstracts.

The results were so impressive (see e.g. Figure 1 and forthcoming article by Johnston, M. and Koentges, T. in Alexandria: The Journal of National and International Library and Information Issues) that the author decided to apply it to Latin and Greek literature in Leipzig's OPP project. OPP has a text collection of over 60 million Greek and Latin words, and has recently begun to add Classical Persian and Arabic texts. It is one of the core interests of OPP to produce methods that can compete with more traditional approaches and that can swiftly generate results on big data. OPP is maintained and organized using eXistDB, the CTS/CITE-Architecture developed by the Homer Multitext Project, and additional web-based tools and services, including GitHub repositories and Morpheus, a Greek and Latin morpho-syntactic analyzer. This structure enables researchers to use a CTS-API to retrieve their desired text-corpora or specific texts. In a first evaluation run of the topic-modeller, 30,000 Classical Arabic biographies have been used (see Figure 2).

In both research institutions, OPP and ATL, researchers applying more qualitative methods complemented the process and evaluated results. In what follows the quantitative method topic-modelling will be explained briefly.

Topic-modelling is "a method for finding and tracing clusters of words (called "topics" in shorthand) in large bodies of texts" (M. Posner, 2012). A topic can be described as a recurring pattern of co-occurring words (M. Brett, 2012). Topic models are probabilistic models that are often based on the number of topics in the corpus being assumed and fixed (D. Bley, 2012). The simplest and probably one of the most frequently applied topic models is the latent Dirichlet allocation (LDA). Success and results of LDA rely on a number of apriori-set variables: for instance, the number of topics assumed in the corpus, the number of iterations of the modelling process, the decision for or against morpho-syntactic normalisation of the research corpus, and how stop-words are implemented in the process. Furthermore, its interpretation is often influenced by how the topics are graphically represented and how the words of each topic are displayed.

While Sievert has found already a very convincing solution for the latter (2014), the former is often out of the hands of the qualitative traditional researcher and any bigger modifications would have to be implemented by a computer-savy researcher. However, topic-modelling is often not an end in itself, rather it is a tool used to help answer a specific humanities research question or to facet


Selected Topic: 20

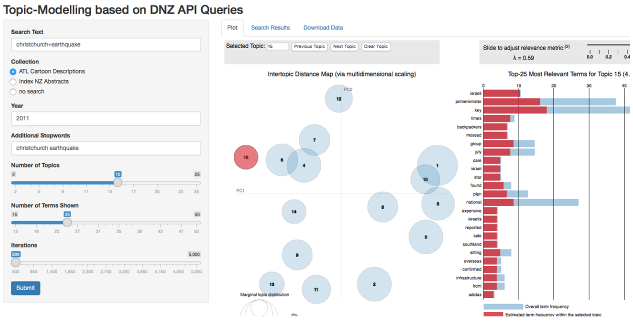
Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 2: LDAvis visualization of the result of an LDA-topic model of 30,000 Classical Arabic biographies, showing a topic that can be used to detect biographies of women.

large text-corpora so that further methods can be applied to a much smaller selection of texts. Traditional researchers often have to continue to work with topic-modelling results, but may not always be aware of the bias that the apriori-set variables have brought into the selection process. One possible way to bridge this gap between researcher and method is to involve the qualitative researcher earlier by providing them with agency in the topic-modelling process.

The author of this paper used R and the web-application framework Shiny to combine Chang's LDA- and Sievert's LDAvis-libraries with DNZ/CTS API requests and language-specific handling of the text data to create a graphical user interface (GUI) that enables researchers to find, topic-model, and browse texts in the collections of ATL, INNZ, OPP, and UMD. They can then export their produced corpus and model, so they can apply qualitative methods on a precise facet of a large text corpora, rather than the whole text corpora itself, which contains texts that are irrelevant for answering the researcher's specific research question. On the left side of the GUI, the researcher can set the following variables: a) search term(s) or CTS-URN(s); b) the source-collection; c) certain stopword lists or processes; d) additional stop-words; e) the number of topics; and f) the number of terms shown for each topic in the visualization. The application then generates the necessary API-requests, normalizes the text as desired by the researcher, applies Chang's LDA-library, and finally presents a D3 visualization of the topics, their relationship to each other, and their terms using Sievert's LDAvis and dimension reduction via Jensen-Shannon Divergence & Principal Components as implemented in LDAvis. The researcher can then directly and visually evaluate the success of their topic-modelling and use the settings on the left like the settings of a microscope to focus on certain significant relationships of word co-occurrences within the corpus. Once they have focused their research tool, they can then export visualizations, topics, and their corpus for further research (see Figure 3) and teaching (see Figure 4).



Topic-Modelling based on DNZ API Queries

Search Text: christchurch+earthquake

Collection
● ATL Cartoon Descriptions
○ Index NZ Abstracts
○ no search

Year: 2011

Additional Stopwords: christchurch earthquake

Number of Topics: 15

Number of Terms Shown: 25

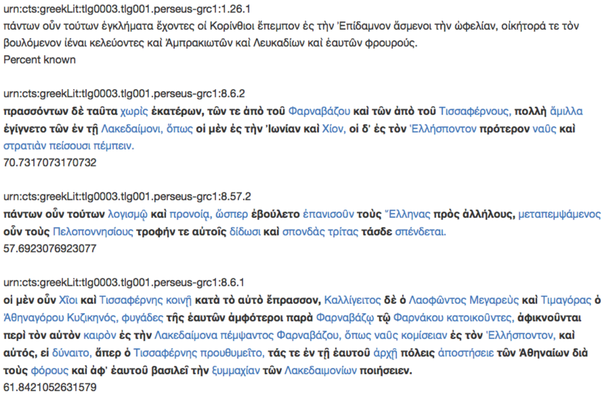
Iterations: 500

Selected Topic: 15

Slide to adjust relevance metric:(2) λ = 0.59

Figure 3: GUI for Topic-Modelling ATL and INNZ descriptions.

The approach also enables educators to use the topic-modelling results to identify sight-readings for language-teaching purposes. The passages are paired based on the profile of their document-topic values and Perseid's Morpheus API is used to mark vocabulary not-yet-known by the learners. Unknown vocabulary links to an online dictionary, using markdown (see Figure 4).



urn:cts:greekLit:tlg0003.tlg001.perseus-grc1:1.26.1
πάντων οὖν τούτων ἐγκλήματα ἔχοντες οἱ Κορίνθιοι ἔπεμπον ἐς τὴν Ἐπίδαμνον ἄσμενοι τὴν ὠφελίαν, οἰκήτορά τε τὸν βουλόμενον ἰέναι κελεύοντες καὶ Ἀμπρακιωτῶν καὶ Λευκαδίων καὶ ἑαυτῶν φρουρούς.
Percent known

urn:cts:greekLit:tlg0003.tlg001.perseus-grc1:8.6.2
πρασσόντων δὲ ταῦτα χωρὶς **ἑκατέρων, τῶν τε ἀπὸ τοῦ** Φαρναβάζου **καὶ τῶν ἀπὸ τοῦ** Τισσαφέρνους, **πολλὴ** ἅμιλλα **ἐγίγνετο τῶν ἐν τῇ** Λακεδαίμονι, ὅπως **οἱ μὲν ἐς τὴν Ἰωνίαν καὶ** Χίον, **οἱ δ' ἐς τὸν** Ἑλλήσποντον **πρότερον** ναῦς **καὶ** στρατιὰν πείσουσι πέμπειν.
70.7317073170732

urn:cts:greekLit:tlg0003.tlg001.perseus-grc1:8.57.2
πάντων οὖν τούτων λογισμῷ **καὶ** προνοίᾳ, ὥσπερ **ἐβούλετο** ἐπανισοῦν **τοὺς** Ἕλληνας **πρὸς ἀλλήλους,** μεταπεμψάμενος **οὖν τοὺς** Πελοποννησίους **τροφήν τε αὐτοῖς** δίδωσι **καὶ** σπονδὰς τρίτας **τάσδε** σπένδεται.
57.6923076923077

urn:cts:greekLit:tlg0003.tlg001.perseus-grc1:8.6.1
οἱ μὲν οὖν Χῖοι **καὶ** Τισσαφέρνης κοινῇ **κατὰ τὸ αὐτὸ ἔπρασσον,** Καλλίγειτος **δὲ ὁ** Λαοφῶντος Μεγαρεὺς **καὶ** Τιμαγόρας **ὁ** Ἀθηναγόρου Κυζικηνός, φυγάδες **τῆς ἑαυτῶν ἀμφότεροι παρὰ** Φαρναβάζῳ **τῷ** Φαρνάκου κατοικοῦντες, **ἀφικνοῦνται περὶ τὸν αὐτὸν** καιρὸν **ἐς τὴν** Λακεδαίμονα πέμψαντος Φαρναβάζου, ὅπως ναῦς κομίσειαν **ἐς τὸν** Ἑλλήσποντον, **καὶ αὐτός, εἰ** δύναιτο, **ἅπερ ὁ** Τισσαφέρνης προυθυμεῖτο, **τάς τε ἐν τῇ ἑαυτοῦ** ἀρχῇ **πόλεις** ἀποστήσειε **τῶν Ἀθηναίων διὰ τοὺς** φόρους **καὶ ἀφ' ἑαυτοῦ βασιλεῖ τὴν** ξυμμαχίαν **τῶν** Λακεδαιμονίων **ποιήσειεν.**
61.8421052631579

Figure 4: Sight-Readings based on the passage Thuc. 1.26.1 with unknown vocabulary marked up.

The author hopes that this paper demonstrates how topic-modelling, in all its complexity, can be opened up to, and positively influenced by, more traditional researchers without advanced computer skills, enabling them to answer specific research questions based on large text corpora.

Bibliography

Blei, D. (2012). "Probabilistic Topic Models." *Communications of the ACM*, **55**(4): 77–84. http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf (accessed 30 October 2015).

Hu, Y., Boyd-Graber J., Satinoff B. and Smith A. (2014). "Interactive topic modeling." *Machine Learning*, **95**(3): 423–69. http://www.umiacs.umd.edu/~jbg/docs/mlj_2013_itm.pdf (accessed 20 February 2015).

Brett, M. (2012). "Topic Modeling: A Basic Introduction." *Journal of Digital Humanities*, **2**(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/ (accessed 30 October 2015).

Posner, M. (2012). "Very Basic Strategies for Interpreting Results from the Topic Modeling Tool." http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/ (accessed 30 October 2015).

Sievert, C. and Shirley, K. (2014). *LDAvis: A Method for Visualizing and Interpreting Topics*, ACL Workshop on Interactive Language Learning, Visualization, and Interfaces. http://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf (accessed 30 October 2015).

The First World War in Perm Provincial Periodicals

Sergei Kornienko
kornienko@psu.ru
Perm State University, Russian Federation

Dinara Gagarina
dinara@psu.ru
Perm State University, Russian Federation

Regional newspapers' periodicals hold a firm place in the humanities among the sources of studying the First World War and its impact on the state and development of the countries and nations in the early 20th century. This kind of sources is especially significant at the level of micro-history and history of everyday life. At the same time, it is complicated to put provincial periodicals into scientific circulation due to problems of access, preservation, organization as well as effective search and analysis of information (Kornienko and Gagarina, 2015).

The source-oriented information system "The First World War in Perm Provincial Periodicals" (http://perm-newspapers.ru/) is an integrated solution to identify problems based on approaches, methods and technologies of Digital Humanities in general and Digital History in particular. The project is realized in the laboratory of historical and political information science of Perm State University in cooperation with the Perm Regional Museum (Russia).

The information system provides free access for scholars to images and full-texts of ten newspapers collections published in Perm province during the First World War. There are more than 2500 issues. Publications cover periods of Imperial Russia, the 1917 Revolution and the Civil War and represent different ideological political movements. These editions include:

- official periodicals of provincial administration of Imperial Russia: "Пермские губернские ведомости" ("Perm Province Bulletin");
- official periodicals of the 1917 Revolution and the Provisional Government period: "Вестник Пермского края" ("Herald of Perm Region"), "Пермский вестник Временного правительства" ("Perm Herald of Provisional Government");
- official newspapers of various levels councils:
- provincial level: "Известия Пермского губернского комитета" ("News of Perm Provincial Committee");branch level: "Известия исполнительного комитета Совета железнодорожных депутатов Пермской железной дороги" ("Proceedings of the Executive Committee of the Board of Railway Deputies of Perm Railroad");district (uyezd) level: "Известия Осинского исполнительного комитета Совета крестьянских, рабочих и солдатских депутатов" ("News of Osa Executive Committee of Peasants, Workers and Soldiers Deputies");
- official gazette of the Perm Province Zemstvo: "Пермская земская неделя" ("Perm Zemstvo Week").

Meaningful diversity and variety of publication's types are distinguishing features of the provincial newspaper periodicals. Official and unofficial newspapers published full texts of normative acts of different levels authorities, official announcements and telegrams, reference and information materials, articles, notes, satires, and others. There were treatments of government, laws and draft laws, regulations, diplomatic notes, information from the fronts of the First World War, reports on meetings of the Government, the Interim Committee of the State Duma, the State Conference and other agencies, including provincial and district authorities, as well as non-government organization. The newspapers published various materials on local history, geography, statistics, ethnography and cover topical issues of socio-economic, political, scientific and cultural life of the country and the region.

Such breadth of perspective and variety of publications leads to a high level of demand for local periodicals as a source of information for humanitarian studies in political, economic and social history, the history of printing and journalism, literature, linguistics, philology, cultural studies, political science, etc.

However, a variety of information submits special claims for its structuring in the information system, the search tools, representation and visualization of electronic versions.

The informational system "The First World War in Perm Provincial Periodicals" is structured based on metadata system that includes thematic fields (rubrics, headings, subheadings of publications), geographical, toponymical fields, personalities. This provides effectiveness of information retrieval and samples formation on various topics, issues, and other criteria at the level of the issue, single newspaper or their combination as well as whole collection. The system also gives possibilities to search keyword and context on titles and full texts for all editions and publications.

The informational system allows visualization of information at the level of newspapers, issues and publications. The results of search requests are displayed in the form of various lists and full-text publications. Each issue of newspapers is presented page by page in PDF format (text below the image), which help to preserve the text content and appearance of newspapers as much as possible. In addition, this method of representation allows text verification and editing as well as reading if OCR is found to be impossible. Texts of publications on the First World War theme are presented in HTML format. Both PDF files and texts as well as all other information and metadata on newspapers, issues and publications are stored in MySQL database.

Organization and structuring of sources data, search tools and visualization tools allow to obtain data on a wide range of issues related to the First World War, life and activities of the region's population in this period.

The information system "The First World War in Perm Provincial Periodicals" provides new possibilities for evaluating potential sources of information, completeness, representativeness and credibility of Perm newspapers periodicals, using computer processing techniques, obtaining new data for scientific humanities research.

Issues that can be studied by the system include attitude towards the war of different social classes and various persons in Perm, the evaluation of forces and the actions of Russia, its allies and opponents. It covers key war events, the role of various commanders, activities of Nicholas II, events in the Perm province and its districts related to the war and their consequences, the creation of images of the war, daily life in the rear and the fronts, etc.

Techniques and methods for solving research tasks are developed based on different types of database queries, which allow obtaining quantitative characteristics, samples for various themes and for specific items of publications. The implementation of these types of queries permits to determine the most common types and genres, subject focus of publications and their relation, generate text fragments and interpretation of the results in terms of the completeness and nature of the information source.

The study is supported by Russian Foundation for Humanities (grant 14-11-59003).

Bibliography

Kornienko, S. and Gagarina, D. (2015). Information systems: new methods of Russian history sources study. *International Multidisciplinary Scientific Conferences Social Sciences AND Arts SGEM 2015. Conference Proceedings. Anthropology, Archaeology, History and Philosophy.* Sofia, pp. 337-43.

Language Attitudes of Twitter Users Toward New York City English

Nathan LaFave
nathan.lafave@nyu.edu
New York University, United States of America

New York City English (NYCE) has long been a stigmatized variety of English. In his seminal research on language use in the New York City dialect, the sociolinguist William Labov referred to New York City as "a great sink of negative prestige" (Labov, 1966)—a characterization that reflected the negative view of NYCE speech shared by non-New Yorkers and New Yorkers alike. Decades later, Preston (2003) elicited extremely low ratings of the New York City dialect on scales of both "correctness" and "pleasantness" by participants from across the US. While these studies present strong evidence of the prevalence of negative language attitudes toward NYCE speech, a more complete picture of linguistic ideology would include what speakers say about NYCE when they are not participating in an academic study. This project seeks to accomplish just that, by examining linguistic ideology with respect to NYCE as espoused by users of the social networking service, Twitter.

Twitter has been recognized as an important resource for humanists and social scientists alike. Scholars have collected and analyzed Twitter messages (tweets) in order to investigate numerous textual and linguistic phenomena such as *lexical variation* (differences in use of synonymous words and phrases, such as *pop* vs. *soda* vs. *coke*). Russ (2012) in particular (see also Bamman, 2011) illustrates the utility of Twitter for examining regionally defined lexical variation through comparison of the geographic distribution of word choices in *geotagged* tweets (with GPS coordinates from which they originated) to more traditionally collected dialectology data. All related research has focused on differences in production. However, I argue that Twitter represents an untapped resource for the investigation of *perceptions* of language use, particularly language attitudes toward regional dialects and differences in their phonetic features (which can be identified by non-standard orthography). Using Twitter solves a primary quandary for language attitude researchers—how to acquire naturally occurring data given the fact that participation in research decreases naturalness.

Tweets containing attitudes and ideology were collected using a range of strategies, including text mining for words—and, crucially, spellings—that reference individual features. To do this, however, it is necessary to determine which features get noted and then which lexical items—and which spellings—are used to signal them. For instance, *cawfee* (also, *cawffee*) is a common orthographic representation of the word "coffee" as pronounced with a raised-THOUGHT vowel, one of the signature dialect features of NYCE. Widely used spellings that reflect *r*-vocalization, another key feature of the NYCE dialect, include *New Yawk* and *fuhgeddaboudit*. In addition to collecting tweets containing orthographic representations of nonstandard features, Twitter search parameters included over 20 terms related to possible names for the dialect itself (e.g., *New York accent, Manhattan dialect, Brooklynese*). These were included in part to determine the extent to which the general public perceives a distinction among speakers from the five boroughs (a distinction which has not been borne out by linguistic analysis).

Repeated automated text mining of Twitter using a Python script to interact with the Twitter API yielded 6,384 tweets that match the aforementioned criteria. Elimination of retweets that did not introduce additional linguistic content and inspection to ensure the tweets reference NYCE produced a final corpus of 1,773 tweets. Relative frequencies of the borough-specific and pan-regional terms in the 1,315 tweets that explicitly reference NYCE by some name reveal that Twitter users most frequently refer to NYCE as the *New York accent* (N=805; 61.2%), though *Brooklyn accent* (N=359; 27.4%) accounts for more than a quarter, with *Bronx accent* (N=54), *Queens accent* (N=29, tied with *Brooklynese*—the most frequent *-ese* moniker), and *Staten Island accent* (N=10) being used much less often. Whether *New York accents* and *Brooklyn accents* are perceived as linguistically or socially distinct, or two names for the same dialect region, will be explored in the paper.

All tweets were manually coded to determine their sentiment with respect to NYCE.

- POSITIVE: *I swear girls from New York accent sound so sexy*
- NEUTRAL: *GAWGEOUS idea she said in her New Yawk accent*
- NEGATIVE: *If you have a Brooklyn accent I automatically want to punch you.*

Almost half of these tweets are neutral in sentiment (N=584, 44.4%); 378 were positive (28.7%) and 200 negative (15.2%). However, 154 tweets were classified as UNCLEAR (8.7%)—many are ambiguous as to whether they evaluate an imitation of an accent or the accent itself, such as when describing an actor's performance (which is common among these types of tweets):

- UNCLEAR: *his New York accent is so bad /:*

Examples such as these pose significant obstacles to automated sentiment analysis—which has been extended to Twitter data (see for instance Pak and Paroubek, 2010)—particularly of language attitudes. Automatic methods would simply code the tweet as negative without recognizing the need to differentiate its underlying meaning. It is noteworthy, however, that even if every UNCLEAR tweet is actually expressing negative sentiment, there would *still* be a greater number of tweets with positive opinions of NYCE speech than negative ones. Furthermore, when Twitter users reference a specific NYCE feature, their evaluation of it is more likely to be positive, regardless of whether they use standard (N=79) or nonstandard (N=568) orthography to represent the feature.

These findings portray a broader range of reactions to NYCE than the language attitudes speakers have presented when engaged in academic research. The paper will include discussion of both negative and positive language attitudes that Twitter users espouse concerning the dialect features associated with NYCE. For instance, any tweets with positive sentiment will be examined to determine if they represent instances of "covert prestige" (Labov 1966), whereby speakers use stigmatized varieties for in-group identification and solidarity. Additional discussion will focus on which regional features evoke the most metacommentary. Furthermore, I will explore the extent to which Twitter users draw (additional) attention to nonstandard forms they employ through capitalization (*NEW YAWK*), hashtags (*#newyawk*), and other orthographic means.

Bibliography

Bamman, D. (2011). Lexicalist. http://www.lexicalist.com/ (accessed 30 August 2015).

Labov, W. (1966). *The Social Stratification of English in New York City*. Washington, D. C.: Center for Applied Linguistics.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of Language Resource and Evaluation Conference* (LREC), Valletta, Malta.

Preston, D. (2003). Language with an attitude. In J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds), *The Handbook of Language Variation and Change*. Oxford: Wiley- Blackwell, pp. 40-66.

Russ, B. (2012). *Examining large-scale regional variation through online geotagged corpora.* Presented at the 2012 American Dialect Society Annual Meeting.

All in the Family: Testing Burrows' Delta on Robert Louis Stevenson's Collaboratively Authored Volumes The Dynamiter and The Wrecker

Anouk Lang
anouk.lang@ed.ac.uk
University of Edinburgh, United Kingdom

Robyn Pritzker
s1259282@sms.ed.ac.uk
University of Edinburgh, United Kingdom

Robert Louis Stevenson (1850-1894) is known to have co-authored several of his works, including *The Dynamiter* (1885), on which he collaborated with his wife Fanny van der Grift Stevenson (Stevenson and Stevenson 1885). Until recently, Stevenson scholars had comparatively little information about how the collaboration around *The Dynamiter* had operated, with one narrative claiming prominence: the preface to the 1905 edition written by Fanny (after her husband's death) which stated that she had invented the stories during an illness of his, and they had subsequently worked collaboratively to put them into written form (Stevenson 1905).

In this paper, we take insights gleaned from our previous work on the authorship of *The Dynamiter* (carried out with collaborators Mingyuan Chen, Carlos Fonseca, Laura McAleese, Alba Morollón Díaz-Faes and Elizabeth

Nicholas) to investigate another text, *The Wrecker* (1892). (At the time of submitting this abstract, the authorship analysis of *The Wrecker* had not yet been carried out, so we have supplied details about our previous study so as to demonstrate that the methods used are both appropriate and robust.) We used the R package Stylo (Eder et al. 2015) to apply Burrows' Delta (Burrows 2002) to two reference corpora containing works known to have been solely authored by Fanny or Robert Louis, and used these as comparators against the individual stories in *The Dynamiter*.

Visualizing the results in the form of a cluster analysis indicated that Burrows' Delta performed well at separating out texts known to be authored by Fanny from texts known to be authored by Robert Louis (Fig 1.)

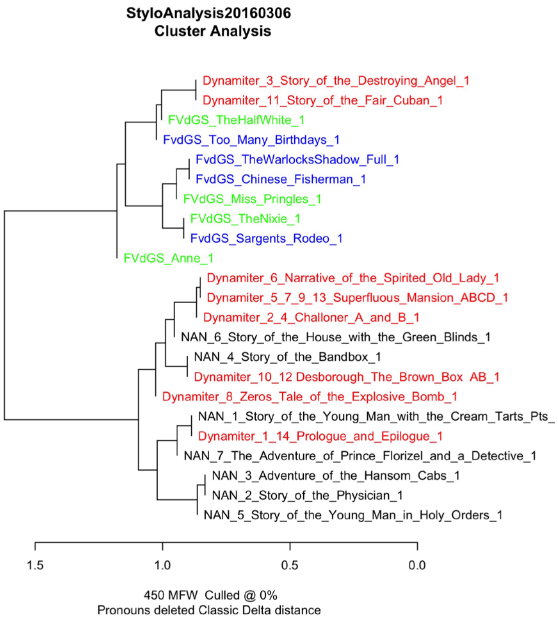


Fig. 1. Cluster analysis of Burrows' Delta scores of the 450 most frequent words in texts by Fanny, in green and blue, by Robert Louis, in orange and black, and *The Dynamiter*, in red (pronouns deleted, no culling).

Our interpretation of these results was that the stories Fanny was most likely to have authored from *The Dynamiter* were "The Story of the Destroying Angel" and "The Fair Cuban". However, it is important to note that while Fanny may well have had an originating role for the plot of many, if not all, of the stories – something which the Preface to the 1905 edition seeks to establish – when it came to the actual writing down of the stories, the "signal" from her linguistic signature was made less clear by the "noise" of her husband's heavy editorial hand (something which is known from biographical and historical writings about their relationship).

Building on this earlier work on *The Dynamiter*, the paper we propose here will examine another work co-authored by Stevenson: *The Wrecker*, also a volume of short stories, which Stevenson co-authored with his stepson Lloyd Osbourne. Stylo will again be used, as will insights from the many *Dynamiter* tests. These indicated, for instance, that deleting pronouns resulted in better separation of texts in the reference corpus, something we attributed to Stevenson's tendency to write about predominantly male characters, which resulted in the prevalence of male pronouns. With the experience gained from attempting to find a solution to the "signal" vs. "noise" problem caused by Robert Louis's proclivity to edit the work of his collaborators, we will also investigate how changes to additional parameters offered by Stylo – changes to the number of most frequent words considered, for example, and variations in text sample size – affect the results of the *Wrecker* tests.

As mentioned above, our earlier results suggested that Robert Louis's editing practices – he had a tendency to edit texts meticulously prior to publication – makes it hard to determine with certainty which sections were written, or initially drafted, by Fanny. A model of co-authorship in which the boundaries between one author and another are clear-cut – where the assumption is that one person is solely responsible for some sections and a second person is solely responsible for others – breaks down in a situation such as this, where the shared domestic space of two authors means that close consultation with – and iterative redrafting of the work of – one's familial collaborator is not simply possible but likely. It underlines the need for stylometric analysis to be complemented with careful literary historical analysis in order to arrive at any meaningful conclusions.

The broader significance of this work is that it is not only of relevance for the field of authorship attribution, but also for its recuperative potential in relation to figures who are less prominent within literary history, including women. Despite scholars' awareness of Fanny's involvement with her husband's writing, for instance, there is still minimal research into how she influenced his work, and how other women in similar positions influenced their famous husbands' legacies. In bringing to light the omitted literary contributions from women and other family members, what emerges is the need for theoretical approaches capable of evaluating the gendered practices of literary studies and book history in contributing to these omissions. In analysing letters and journals, combing through stylometric data, and assessing biographical accounts, there is a risk of overlooking collaborations in favour of other types of partnerships (amanuenses, muses, and the like). A woman like Fanny van der Grift, who history has recorded as a keen writer, diarist, and editor, surpasses the labels that book history has offered her. While authorship attribution analysis offers one set of useful tools for breaking down these barriers, further critical engagement with our own gendered scholarly practices is necessary to more clearly understand how and why certain works and canons have become established in the way that they have.

Bibliography

Burrows, J. F. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship, *Literary and Linguistic Computing*, **17**(3): 267-87.

Eder, M., Rybicki, J. and Kestemont, M. (2015). 'Stylo': A Package for Stylometric Analyses. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxjb21wdXRhdGlvbmFsc3R5bGlzdGljc3xneDpmM2U3OGUzZTM2YjkyYzM (accessed 6 March 2016).

Stevenson, F.V.d.G. (1905). Preface to the Biographical Edition. In Stevenson, R. L. and Stevenson, F.V.d.G., *More New Arabian Nights: The Dynamiter*. New York: Scribners, pp. V–XIV.

Stevenson, R. L. and Stevenson, F.V.d.G. (1885). *The Dynamiter: More New Arabian Nights*. London: Longmans, Green, and Co.

Morphology beyond inflection. Building a wordformation based dictionary for Latin

Eleonora Litta
e.littamodignani@gmail.com
Universita' Cattolica del Sacro Cuore, Italy

Marco Carlo Passarotti
marco.passarotti@unicatt.it
Universita' Cattolica del Sacro Cuore, Italy

This short paper describes the initial phases of a Marie Curie Research Project, *Word Formation Latin* (WFL), developed at the Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE), at the Università Cattolica del Sacro Cuore, Milan, Italy. The project consists in the compilation of a derivational morphological dictionary of the Latin language, which connects lexical elements on the basis of wordformation rules, through the use of computational linguistic methods.

In the past two decades there has been a considerable increase in the creation of computational language resources for the investigation of classical languages, which have updated the state of the art almost to the same level as that of the resources currently available for modern languages.

However, among the existing language resources, we currently lack, for Latin, a morphological derivational dictionary that connects lexical elements on the basis of Word Formation Rules[1] (WFRs).

A first attempt at constructing a lexicon based on wordformation for Latin was made by Marco Passarotti and Francesco Mambrini in 2012 (Passarotti & Mambrini, 2012). The WFL project has been awarded funding to expand on these efforts.

The project has three main aims:

1. the enrichment of an existing morphological analyser for the Latin language, LEMLAT (Passarotti, 2004), with wordformation information, and the integration of data within a interface similar to Word Manager (Domenig & ten Hacken, 1992), which has been already applied to other modern languages (English, German, Italian);

2. the integration of the information extracted from the resulting derivational morphological dictionary into the morphological layer of annotation the *Index Thomisticus* Treebank (IT-TB);[2]

3. offering the results of the project work via a user-friendly project website which will display the derivational morphological dictionary through a web based search interface. This will allow the lexicon to be accessed:

   1. by single lexical entry, which will show both the ancestors and their derived words;
   2. by morphological family;[3]
   3. by WFR.

The project relies on the automatic realisation of the linguistic resource both at the level of WFRs creation and to their application on the lexical items included in the morphological analyser LEMLAT. The LEMLAT lexical basis contains around 40.000 lemmas from three major Latin dictionaries (Georges, 1913-1918; Gradenwitz, 1904; Glare, 1982). We conceived WFRs according to the so-called Item-and-Arrangement model (IA), which follows a morpheme-based approach to morphology. In IA, word forms are analysed as arrangement of morphemes according to the following three axioms:

1. Roots and affixes have the same status of morphemes (Baudoin's single morpheme hypothesis);

2. They are dualistic, as they have both a form and a meaning (Bloomfield's sign base morpheme hypothesis);

3. They are stored in the lexicon (Bloomfield's lexical morpheme hypothesis).

The aim is to assign a WFR to each morphologically complex lemma (i.e. one morphologically derived from another lemma) and to link each complex lemma to its ancestor. The data are organised and presented according to a system similar to that for morphological dictionaries devised by Word Manager, in which relations between the members of the same morphological family are represented in a tree-graph.

WFRs are grouped in two classes: 1. compounding; 2. derivational. Derivational rules are divided in two categories: a. affixal (in its turn split into prefixal and suffixal), and b. conversive, a derivation process that does not imply any affix; these are manually defined.

This happens in two steps:

1. Phase A: Semi-automatic data-driven finding of WFRs:

1. lemmas are divided into two classes, according to their part of speech and inflectional category;

2. two lists are produced: an *incipitarium* and an

explicitarium, where lemmas are ordered according to a traditional alphabetical order, or to a right-to-left alphabetical order respectively;

3. prefixal and suffixal rules are created from the two lists respectively, part of speech and inflectional category of the lemma(s) are manually assigned to each candidate rule.

2. Phase B: Application (and evaluation) of the WFRs resulting from Phase A, and creation of the "morphological families". New rules are added in this phase by confrontation with data. Phase B is divided into two subtasks:

1. each complex lemma is assigned a WFR. This task is performed by assigning in semi-automatic fashion to each (possibly) complex lemma its most likely WFR according to the PoS of the lemma and the string of its initial (prefixal rules) and final (suffixal rules) characters;
2. morphological families are built.

All those (morphologically simple, or complex[4]) lemmas that share the same invariable part are automatically assigned to the same morphological family.

Finally, the members of each family are automatically linked to each other according to their PoS, inflectional category and affixes by means of the WFR assignment (2.a). The morphologically simple (i.e. not derived) lemma member is assigned the role of ancestor of the family.

Phase A finds the WFR, Phase B applies the WFR to data, obtaining input and output lemmas for each WFR.

Phase A is not to be considered exhaustive, but exploratory: the recall of WFR identified in Phase A is not 100%. The aim in the first phase of the project is to refine the data by tagging the highest number of lexemes using data driven WFRs, which will be increasingly complex, covering most well known wordformation issues.[5] Given the high number of homographs in Latin, this automatic procedure is regarded as non-ultimate for building the morphological families. However, it is helpful as it provides filtered data that must be checked manually.

This is why we need Phase B during which, by comparison with the evidence given by data, we can identify the rules that were missed in phase A. Manual hardcoding will be necessary for those lemmas produced by poorly productive WFRs, or morphotactically obscure wordformation processes. Evaluation of the language resource is performed by manual checking data organised into homogeneous groups based on WFRs (coverage of rules) and stemming (coverage of morphological families). Precision and recall are used as evaluation metrics in order to calculate the rate of positive and negative cases.

To date, 118 WFRs have been found automatically. Around 50 of these rules, those showing a certain degree of morphological transparency, hence easier to obtain through the automatic finding in the input-output relation (e.g. derivational, verb-to-verb, prefixal, etc.), have been added to a SQL database, and resulted in the tagging of some 9000 morphologically complex lexemes.

The final resource will be both a standalone dictionary accessible through its own website, and interconnected with the *Index Thomisticus*.

The integration with the IT-TB will be operated through the embedding of the dictionary data within the morphological layer of annotation of the treebank, using TEI (Text Encoding Initiative) P5 conformant XML encoding to favour data exchange and linking to other lexical resources. The data resulting from the dictionary, once encoded in XML, will be applied to the IT-TB data.

Bibliography

Busa, R. (1988). *Totius latinitatis lemmata*. Milano: Istituto Lombardo - Accademia di Scienze e Lettere.

Domenig, M. & ten Hacken, P. (1992). *Word Manager: A system for morphological dictionaries*. Hildesheim: Olms.

Forcellini, A. (1771). *Totius Latinitatis lexicon, consilio et cura Jacobi Facciolati opera et studio Aegidii Forcellini, lucubratum*. Patavii: typis Seminarii, 4 voll.

Georges, K. E. (1913-1918). *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hannover: Hahn.

Glare, P. G. W. (1982). *Oxford Latin Dictionary*. Oxford.

Gradenwitz, O. (1904). *Laterculi Vocum Latinarum*. Leipzig.

Index Thomisticus Treebank. http://itreebank.marginalia.it/.

LEMLAT. http://www.ilc.cnr.it/lemlat/lemlat/index.html.

Lewis, C. T. and Short, C. (1969). *A Latin Dictionary*. Oxford: At the Clarendon press.

McGillivray, B. and Passarotti, M. (2009). "The Development of the Index Thomisticus Treebank Valency Lexicon". In *Proceedings of LaTeCH-SHELT&R Workshop 2009*. Athens, March 30.

Passarotti, M. (2004). "Development and perspectives of the Latin morphological analyser LEMLAT". In A. Bozzi, L. Cignoni and J. L. Lebrave (Eds.), *Digital Technology and Philological Disciplines. Linguistica Computazionale*, XX-XXI, pp. 397-414.

Passarotti, M. and Mambrini, F. (2012). "First Steps towards the Semi-automatic Development of a wordformation-based Lexicon of Latin", In *Proceedings of LREC 2012*. Istanbul, Turkey, pp. 852-59.

TEI Consortium, (eds). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 2.9.1. 2015-10-15. TEI Consortium. http://www.tei-c.org/Vault/P5/current/ (accessed 22 February 2016).

Notes

[1] Word formation is the creation of a new word from either the combination of two other words (*dish-washer*, compounding) or of adding one of more affixes to an existing word (*wash-er*, derivation), or from a part of speech change (*clean*, verb vs. *clean*, adjective).

[2] The *Index Thomisticus* (IT) is considered a pathfinder in digital humanities; started by Padre Roberto Busa in 1949. It is a database retaining the *opera omnia* by Thomas Aquinas (118 texts), plus works by other 61 authors related to Thomas

(61 texts). The size of the corpus is around 11 million tokens (150.000 types; 20.000 lemmas). The corpus is fully lemmatised and morphologically tagged. The IT-TB, based at CIRCSE, is the syntactically annotated portion of the IT, and it contains around 300.000 tokens for 15.000 syntactically parsed sentences.

[3] By "morphological family" we mean the set of lemmas morphologically derived from one common ancestor-lemma

[4] WFRs do not take in input morphologically simple lemmas only, but also complex ones. For example, the noun *excubatio* derives by suffixation from the verb *excubo*, which is morphologically complex, as it is derived (by prefixation) from the verb *cubo*.

[5] i.e. stem change featuring internal vowel alternation (*fac.io*, *per-fic-io*), assimilation of prefix (*fer-o* > **ob-fer-o* > *of-fer-o*), unclear segmentation (*cre-a-tor* or *cre-at-or*?), etc.

Analyzing the 17th Century Theatre Critique Texts with a Semantic Annotation Tool Driven by a Dedicated Ontology

Chiara Mainardi
chiara85.mc@gmail.com
Université Paris-Sorbonne, France; Université Pierre et Marie Curie, France

Vincent Jolivet
vincent.jolivet@paris-sorbonne.fr
Université Paris-Sorbonne, France; Université Pierre et Marie Curie, France

Zied Sellami
contactzied@gmail.com
Université Paris-Sorbonne, France; Université Pierre et Marie Curie, France

Project "Haine du Théâtre"

This experiment takes place within the "Haine du Théâtre"[1] Project, which aims at analysing theatre debates in Europe by using scientific approaches and critical editions of polemical texts. The reflections of the team were primarily focused on the discovery of the circumstances and arguments used in theatre all across Europe, not limited to France, but including England, Spain, Italy, and the Germanic area. The timeframe encompassed the last decades of the 16th century up to the beginning of the 19th century. The purpose of the Project is to explore the grey areas of the controversies in order to outline a global overview of the situations which led to these polemics, discovering where and how they began, their chronological discrepancies in the different countries, and the links between them and their contemporary resurgences.

Corpus

The total collection of the Haine du Théâtre Project related to France made up of 300 texts in the PDF format. The XML/TEI critical edition of 27 texts has been achieved manually (examples of the main titles of this TEI corpus are D'Aubignac (1666), Conti (1666), Pierre Nicole (1667), Voisin (1671), Vincent (1647), cfr. <http://obvil.paris-sorbonne.fr/corpus/haine-theatre/>) and this small corpus is used for the semantic analysis we present here.

This collection is a homogenous combination of texts in which the different authors express their approval or their condemnation of theatre. Using the ontology we created, our interest was to discover to what extent the personal judgment of each author was celebratory or derogatory about theatre and, also, which arguments were mostly involved in their critiques.

In this direction, we had two main goals:

1. Organizing the knowledge about polemics in theatre and their vocabulary;

2. Use this structured lexicon for the corpus annotation and its analysis.

Building of the ontology

An ontology is a good way to automatically analyse many texts together along two parameters: documentary and linguistic. On the one hand, by improving a state of knowledge about 17th century French ; on the other hand, by refining the vocabulary linked to theatre controversy.

The ontology organizes the knowledge of our domain (polemical texts about theatre) as structured points of view (condemnation, defence, etc.) as well as 44 structured classes (concepts) related to critical controversies.

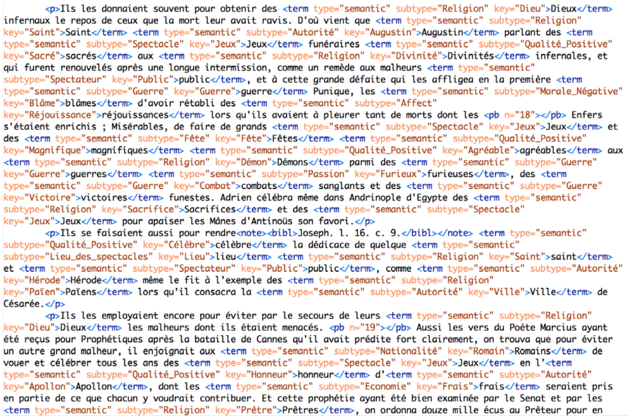
These classes report the axiological points of view (the judgment of the authors), their objects (such as *jongleur*, *actrice*, etc.), and the thematics of the polemics (*religion*, *emotions*, etc.).

To detect the salient terms pertaining to each concept, we realized that in most texts of the modern period, authors' judgements revolve around quoted authorities. The context and the deep knowledge of the corpus together with expertise in 17th century French constitute the cornerstones of our approach. It is very important to fully understand the contextual meaning (in the 17th century) of the selected salient terms.

The idea we came up was to look at the relative importance of the various semantic fields in each chapter of the collection, then extract a priori the content of a chosen text. We began to use this structured lexicon of outstanding terms related to theatre polemics for the corpus annotation and its analysis.

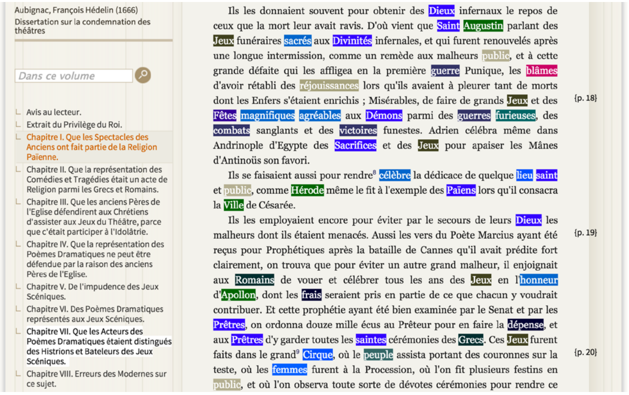
To our knowledge no other comparable ontology about theatre polemics exists; therefore, we will present the method we conceived keeping in mind that those classes are not exhaustive and that new questions will require the creation of new classes.

Despite the high specificity of the domain, this ontology model and the automatic annotator are deployable in other contexts, such as literary criticism and theatre critique in different languages. For this purpose we will translate the model into English (the set of the attributes useful for the annotator).



```
        <p>Ils les donnaient souvent pour obtenir des <term type="semantic" subtype="Religion" key="Dieu">Dieux</term>
infernaux le repos de ceux que la mort leur avait ravis. D'où vient que <term type="semantic" subtype="Religion"
key="Saint">Saint</term> <term type="semantic" subtype="Autorité" key="Augustin">Augustin</term> parlant des <term
type="semantic" subtype="Spectacle" key="Jeux">Jeux</term> funéraires <term type="semantic" subtype="Qualité_Positive"
key="Sacré">sacrés</term> aux <term type="semantic" subtype="Religion" key="Divinité">Divinités</term> infernales, et
qui furent renouvelés après une longue intermission, comme un remède aux malheurs <term type="semantic"
subtype="Spectateur" key="Public">public</term>, et à cette grande défaite qui les affligea en la première <term
type="semantic" subtype="Guerre" key="Guerre">guerre</term> Punique, les <term type="semantic" subtype="Morale_Négative"
key="Blâme">blâmes</term> d'avoir rétabli des <term type="semantic" subtype="Affect"
key="Réjouissance">réjouissances</term> lors qu'ils avaient à pleurer tant de morts dont les <pb n="18"></pb> Enfers
s'étaient enrichis ; Misérables, de faire de grands <term type="semantic" subtype="Spectacle" key="Jeux">Jeux</term> et
des <term type="semantic" subtype="Fête" key="Fête">Fêtes</term> <term type="semantic" subtype="Qualité_Positive"
key="Magnifique">magnifiques</term> <term type="semantic" subtype="Qualité_Positive" key="Agréable">agréables</term> aux
<term type="semantic" subtype="Religion" key="Démon">Démons</term> parmi des <term type="semantic" subtype="Guerre"
key="Guerre">guerres</term> <term type="semantic" subtype="Passion" key="Furieux">furieuses</term>, des <term
type="semantic" subtype="Guerre" key="Combat">combats</term> sanglants et des <term type="semantic" subtype="Guerre"
key="Victoire">victoires</term> funestes. Adrien célébra même dans Andrinople d'Egypte des <term type="semantic"
subtype="Religion" key="Sacrifice">Sacrifices</term> et des <term type="semantic" subtype="Spectacle"
key="Jeux">Jeux</term> pour apaiser les Mânes d'Antinoüs son favori.</p>
        <p>Ils se faisaient aussi pour rendre<note><bibl>Joseph. l. 16. c. 9.</bibl></note> <term type="semantic"
subtype="Qualité_Positive" key="Célèbre">célèbre</term> la dédicace de quelque <term type="semantic"
subtype="Lieu_des_spectacles" key="Lieu">lieu</term> <term type="semantic" subtype="Religion" key="Saint">saint</term>
et <term type="semantic" subtype="Spectateur" key="Public">public</term>, comme <term type="semantic" subtype="Autorité"
key="Hérode">Hérode</term> même le fit à l'exemple des <term type="semantic" subtype="Religion"
key="Païen">Païens</term> lors qu'il consacra la <term type="semantic" subtype="Autorité" key="Ville">Ville</term> de
Césarée.</p>
        <p>Ils les employaient encore pour éviter par le secours de leurs <term type="semantic" subtype="Religion"
key="Dieu">Dieux</term> les malheurs dont ils étaient menacés. <pb n="19"></pb> Aussi les vers du Poète Marcius ayant
été reçus pour Prophétiques après la bataille de Cannes qu'il avait prédite fort clairement, on trouva que pour éviter
un autre grand malheur, il enjoignait aux <term type="semantic" subtype="Nationalité" key="Romain">Romains</term> de
vouer et célébrer tous les ans des <term type="semantic" subtype="Spectacle" key="Jeux">Jeux</term> en l'<term
type="semantic" subtype="Qualité_Positive" key="Honneur">honneur</term> d'<term type="semantic" subtype="Autorité"
key="Apollon">Apollon</term>, dont les <term type="semantic" subtype="Economie" key="Frais">frais</term> seraient pris
en partie de ce que chacun y voudrait contribuer. Et cette prophétie ayant été bien examinée par le Senat et par les
<term type="semantic" subtype="Religion" key="Prêtre">Prêtres</term>, on ordonna douze mille écus au Prêteur pour en
```

Snapshot of the TEI annotation (François Hédelin d'Aubignac, *Dissertation sur la condemnation des théâtres*, 1666, "Chapitre I. Que les Spectacles des Anciens ont fait partie de la Religion Païenne")



Aubignac, François Hédelin (1666)
Dissertation sur la condemnation des théâtres

- Avis au lecteur.
- Extrait du Privilège du Roi.
- Chapitre I. Que les Spectacles des Anciens ont fait partie de la Religion Païenne.
- Chapitre II. Que la représentation des Comédies et Tragédies était un acte de Religion parmi les Grecs et Romains.
- Chapitre III. Que les anciens Pères de l'Eglise défendirent aux Chrétiens d'assister aux Jeux du Théâtre, parce que c'était participer à l'Idolâtrie.
- Chapitre IV. Que la représentation des Poèmes Dramatiques ne peut être défendue par la raison des anciens Pères de l'Eglise.
- Chapitre V. De l'impudence des Jeux Scéniques.
- Chapitre VI. Des Poèmes Dramatiques représentés aux Jeux Scéniques.
- Chapitre VII. Que les Acteurs des Poèmes Dramatiques étaient distingués des Histrions et Bateleurs des Jeux Scéniques.
- Chapitre VIII. Erreurs des Modernes sur ce sujet.

Ils les donnaient souvent pour obtenir des Dieux infernaux le repos de ceux que la mort leur avait ravis. D'où vient que Saint Augustin parlant des Jeux funéraires sacrés aux Divinités infernales, et qui furent renouvelés après une longue intermission, comme un remède aux malheurs public, et à cette grande défaite qui les affligea en la première guerre Punique, les blâmes d'avoir rétabli des réjouissances lors qu'ils avaient à pleurer tant de morts dont les Enfers s'étaient enrichis ; Misérables, de faire de grands Jeux et des {p. 18} Fêtes magnifiques agréables aux Démons parmi des guerres furieuses, des combats sanglants et des victoires funestes. Adrien célébra même dans Andrinople d'Egypte des Sacrifices et des Jeux pour apaiser les Mânes d'Antinoüs son favori.

Ils se faisaient aussi pour rendre[8] célèbre la dédicace de quelque lieu saint et public, comme Hérode même le fit à l'exemple des Païens lors qu'il consacra la Ville de Césarée.

Ils les employaient encore pour éviter par le secours de leurs Dieux les malheurs dont ils étaient menacés. Aussi les vers du Poète Marcius ayant été {p. 19} reçus pour Prophétiques après la bataille de Cannes qu'il avait prédite fort clairement, on trouva que pour éviter un autre grand malheur, il enjoignait aux Romains de vouer et célébrer tous les ans des Jeux en l'honneur d'Apollon, dont les frais seraient pris en partie de ce que chacun y voudrait contribuer. Et cette prophétie ayant été bien examinée par le Senat et par les Prêtres, on ordonna douze mille écus au Prêteur pour en faire la dépense, et aux Prêtres d'y garder toutes les saintes cérémonies des Grecs. Ces Jeux furent faits dans le grand[9] Cirque, où le peuple assista portant des couronnes sur la {p. 20} teste, où les femmes furent à la Procession, où l'on fit plusieurs festins en public, et où l'on observa toute sorte de dévotes cérémonies pour rendre ce

Snapshot of the HTML highlighting of the annotation (François Hédelin d'Aubignac, *Dissertation sur la condamnation des théâtres*, 1666, "Chapitre I. Que les Spectacles des Anciens ont fait partie de la Religion Païenne"). The salient words of the religion thematic, very important in this chapter, are highlighted in dark blue ("Dieux", "Saint", "Divinités", "Démons", "Sacrifices", "Païens", "Dieux", "Prêtres", etc.)

The annotation tool

We created an annotator in order to markup all the forms of the outstanding terms, recorded as lemmas in the ontology ("horrible", "horribles", etc.), except when:

- the exact form matters: the tool annotates only the exact form ("Père" does not match "père", "pères");
- the gender matters: only the feminine forms, singular and plural ("courtisane(s)" which is a linguistic sign of the "Femme" concept in our ontology does not match the masculine forms "courtisan(s)").

The annotation tool enriches automatically the TEI files by setting down some <term> tags to mark all the forms, but also their lemma and the related semantic field.

Following this method we find a high density of salient terms (on average 9,5% of words in each the chapters) in our collection of texts. The combination of these linguistic signs and related semantic fields emerges as a semantic descriptor of the corpus content (e.g. http://obvil-dev.paris-sorbonne.fr/corpus/haine-theatre/vincent_traite-des-theatres_1647/vincent_traite-des-theatres_1647_6).

Semantic analysis

Without preconceived notions and previous knowledge of the corpus, it was possible for us to answer numerous research questions. By examining the intensity of condemnation in the corpus, we discovered that the most derogatory chapters belong to Nicole and Conti.

chapter_id	condemnation_ terms	defense_ terms	chapter_ length (words)	condemnation intensity (‰)
nicole_traite-de-la-comedie_1667_4	6	0	762	7,87
nicole_traite-de-la-comedie_1675_4	6	0	906	6,62
conti_traite-de-la-comedie_1666_14	21	1	4536	4,63

Examples of the condemnation intensity

By virtue of the possibility of making cross queries across multiple classes, we were able to obtain a list of the arguments developed by the authors – which can be confirmed by reading the chapters concerned. For instance, the most derogatory chapters about theatre were written by Nicole and Conti.

chapter_id	"Théâtre" and "Morale Négative" thematics relative stores 2(‰)
nicole_traite-de-la-comedie_1667_4	66,67
nicole_traite-de-la-comedie_1675_4	56,25

nicole_traite-de-la-comedie_1675_1	41,67
conti_traite-de-la-comedie_1666_14	41,24
nicole_traite-de-la-comedie_1675_13	40,75

The most derogative chapters

We can also compare the chapters more concerned with, for instance, condemnation and the thematic of "Women". We discovered that the chapters more concerned with misogyny belonged to Voisin, Conti and Nicole. The discriminant terms for woman are, in these derogatory chapters: "femme, fille, bouffonne, maîtresse, comédienne". Those are just examples of the queries we can construct with the annotation tool.

At this point, the results of the computational analysis reveal some critical conclusions about the textual tradition analysed. Continuing this kind of analysis for every concept present in the ontology, we found that some authors are more concerned with all the elements of theatrical debates, whereas others deal only with some of them.

In particular, we found a high concentration of annotated terms in Voisin, Conti, Nicole and Aubignac, whose chapters usually occupy the first 20 results for most thematics. On the contrary, authors like Vincent, Guillot-Gorju, Le Marcant and Gaule score high only when relating to economic issues.

chapter_id	**"Economy" thematic relative stores[3](‰)**
vincent_traite-des-theatres_1647_9	3.77
vincent_traite-des-theatres_1647_7	3.40
anonyme_honneur-theatre_1620_1	3.1
guillot-gorju_apologie_1634_1	2.75
lemarcant_conduite-du-vrai-chretien_1694_1	2.50
nicole_traite-de-la-comedie_1667_18	2.33

The concentration of the economic thematic. These results show that the economy is at the core of the descriptions by another group of authors: Vincent, Guillot-Gorju, Le Marcant.

For example, some very specific arguments are associated in the controversy, like women, passion and the economy. Moreover, results have shown that some authors of the corpus focus on only few topics of the polemics, whereas others, like Vincent, concentrate all the topics of the theatre polemic in his 11th chapter.

Conclusions

The HdT ontology organizes the knowledge about the theatre polemics and their vocabulary. In particular, it presents a hierarchical lexicon of salient terms and it reflects the points of view, objects and thematics of the polemic.

This ontology enables the exploration the corpus for research questions: the relative frequency of the salient terms (*lemma*) and of their related class (*catégorie d'indexation*) is a semantic descriptor of each chapter. Those descriptors can be exploited for the semantic exploration of the corpus, capitalizing on the ontology structure (terms / classes). The researchers can understand the intensity of the concepts they choose to analyse and thus can focus on the most significant chapters of the corpus.

Bibliography

Aubignac, abbé d'. (1666). *Dissertation sur la condemnation des théâtres*. Paris: N. Pépingué.

Buitelaar, P., Cimiano, P. and Magnini, B. (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*. Frontiers in Artificial Intelligence and Applications Series. Amsterdam: IOS Press.

Cimiano, P., Buitelaar, P. and Volker, J. (2010). Ontology construction. *Handbook of Natural Language Processing*, Second Edition. CRC Press, Taylor and Francis Group, pp. 577–604.

Conti, Prince de. (1666). *Traité de la Comédie et des spectacles*. Paris: L. Billaine.

Gaule, André de. (1607). *Conviction*. Lyon: A. Cloquemin.

Guillot-Gorju, Harduin de Saint-Jacques, Bertrand. (1634). *Apologie de Guillot-Gorju. Addressee à tous les beaux Esprits*. Paris: M. Blageart.

Le Marcant, J. (1694). *La conduite du vrai chrétien*. Paris: E. Couterot.

Maedche, A. (2002). *Ontology learning for the Semantic Web*. Volume 665. Kluwer Academic Publisher.

Nicole, P. (1667). *De la Comédie*. Liege: Adolphe Beyers.

Vincent, P. (1647). *Traité des théâtres*. La Rochelle: J. Chuppin.

Voisin, Joseph de, abbé. (1671). *Défense du traité de Mgr le Prince de Conti touchant la comédie et les spectacles ou la réfutation d'un livre intitulé Dissertation sur la condamnation des théâtres*. Paris: Coignard.

Notes

[1] The directors of this Projet are François Lecercle and Clotilde Thouret. <http://obvil.paris-sorbonne.fr/projets/la-haine-du-theatre> It is one of the many outstanding projects at the Labex OBVIL (Laboratoire d'Excellence: Observatoire de la Vie Littéraire) in Paris headed by Didier Alexandre and Jean-Gabriel Ganascia.

[2] The score is the ratio between the theatre and the condemnation terms over the chapter length (number of words).

[3] The score is the ratio between the economy terms and the chapter length (number of words).

Méthodes computationnelles et analyse d'une langue de chancellerie: le logiciel d'analyse textuelle Machiato et la correspondance diplomatique et administrative de Machiavel

Corinne Manchio
corinne.manchio@gmail.com
Paris 8, France

Marc Lasson
marc.lasson@gmail.com
no affiliation

Cet exposé a pour vocation de présenter un projet en cours qui associe les méthodes traditionnelles des sciences humaines (philologie, histoire et histoire des idées) aux approches computationnelles et statistiques. La création et la mise en place du logiciel d'analyse textuelle MACHIATO (Lasson et Manchio, 2015) constitue le point de convergence de deux perspectives de recherche initialement séparées, celle d'un chercheur en informatique spécialiste de la formalisation mathématique et d'une doctorante en études italiennes, travaillant sur la correspondance diplomatique et administrative de Machiavel (*Legazioni e Commissarie, Scritti di governo*, 2002-2012). La mise en place du site internet relève donc en premier lieu d'une coopération, terme central pour tout chercheur en sciences humaines et sociales qui entreprend de se lancer dans un projet relevant des humanités numériques. En effet, la rencontre entre des champs disciplinaires, implique un constant pour créer un langage commun pour appréhender les textes, pour identifier le type de questionnements qu'il est possible de soumettre aux machines et pour dégager les potentialités et les limites de certaines procédures d'automatisation.

1. Question de méthodes

Nous avons suivi un cheminement complexe depuis la philologie traditionnelle, entendue comme combinaison de critique littéraire, historique et linguistique ou comme pratique d'établissement de texte à partir de sources différentes jusqu'à la philologie dite numérique. L'école qui a le plus influencé nos travaux est celle de Jean-Claude Zancarini et Jean-Louis Fournel et leur approche de ce qu'ils appellent « philologie politique ». Ils se sont attachés à historiciser la pensée de Machiavel, et plus généralement cette génération de la guerre, à travers une analyse minutieuse de son usage des mots dans le cadre de la traduction des textes politiques majeurs de la période des guerres d'Italie. Or, le fait de revenir à chaque usage pour parvenir à une élucidation sémantique est d'autant plus ardu que le corpus est important : pour exploiter nos 2214 lettres, nous avons donc rapidement pensé à croiser notre méthode initiale avec les potentialités des méthodes computationnelles et statistiques, qui nous ont ainsi permis de multiplier les points de vue sur le texte (diachronique, synchronique), de modifier l'unité de base de l'analyse (lemme, champ lexical, champ sémantique) et dégager les tendances générales du corpus et les spécificités de sous-corpus. De telles possibilités modifient inévitablement le rapport aux textes du chercheur en SHS qui ne peut parvenir à l'interprétation des résultats qu'à condition d'en comprendre le sens et les enjeux et donc, de suivre une formation (notamment en statistique et en visualisation de données).

2. Les mots et les nombres: MACHIATO

Le logiciel est commandé par une interface Web qui confère à l'utilisateur un accès facile aux ressources du programme avec son navigateur. Les analyses et les précalculs statistiques sont faits par un back-end mis en œuvre avec le langage de programmation Python dont le rôle est d'initialiser la base de données. Nous utilisons le Framework *open-source* Django pour générer du contenu HTML. Enfin, nous utilisons le langage de programmation Javascript pour afficher les données et les résultats côté utilisateur. Le corpus a tout d'abord été normalisé : différents types d'entrée permettent de le découper en fonction de nos besoins. L'index des lettres permet d'accéder à chaque missive en obtenant les indications de base sur chacune d'elles ; l'index des occurrences donne accès à chacune des 25 007 formes, permettant de cibler des graphies particulières, des noms propres ou des hapax) ; l'index des 6399 familles de mots (qui rassemble toutes les flexions d'un vocable, très nombreuses pour certains verbes du fait de leur emploi et de l'instabilité graphique très forte) consent de mesurer le poids d'un champ lexical. En outre, à chaque missive ont été associés son destinataire, sa date et son lieu de rédaction. Ce *tagging* préliminaires nous a ensuite donné la possibilité de formuler de nouvelles questions plus complexes et de croiser différentes variables (à titre d'exemple, quel terme est le plus employé lorsque Machiavel se trouve en mission auprès du roi Louis XII ?).

Nous avons repris une large part des outils proposés dans les logiciels d'analyse textuelle existants et utilisés par les linguistes à l'instar des concordances, particulièrement utiles pour dégager rapidement les nombreux homographes de notre corpus, et des cooccurrences indispensables pour repérer les cas de polysémie, voire de double isotopie typiques de la langue machiavélienne. Nous avons en outre emprunté un modèle d'analyse des cooccurrences multiples (Heiden, 2004) et les principes de mesure et de comparaison développés en statistique textuelle par Lebart et Salem. Les différents types de calcul permettent de décrire de façon toujours plus précise non seulement la nature des usages, mais aussi leur poids et les relations qui les lient les unités du texte entre elles. Nous utilisons trois types de calcul : les fréquences, l'indice de dispersion et l'indice de spécificité qui ont en outre permis de dépasser certains reproches récurrents faits aux méthodes computationnelles (dont celui de ne considérer la langue comme un ensemble d'unités abstraites, sans lien et décontextualisées). Nous avons fait le choix de limiter les types de visualisation pour des raisons de cohérence, de temps et de compétence. Nous proposons d'en donner quelques exemples tels que l'histogramme de répartition des fréquences, celui des graphies, les structures arborescentes (permettant de représenter tous les environnements immédiats en une image) et les histogrammes interactifs représentant la chronologie des missives. Toutes les visualisations de données sont exportables : nous utilisons Excel pour les tables, ainsi que le format SVG pour les données graphiques. Les critères de recherches avancées (permettent d'effectuer des recherches complexes en fonction du nombre d'occurrences ou/et de la fréquence et/ou des indices de répartition et de dispersion). À partir de l'index des lettres, il est aussi possible de questionner un sous-corpus particulier en vertu de critères variés tels que les bornes chronologiques ou la nature des missives (lieu de rédaction). Il est aussi possible de sélectionner un sous-corpus spécifiques et de procéder à une recherche précise en activant (ou non) les filtres sur les occurrences suivants : taille, nombre d'occurrences dans le sous-corpus interrogé, nombre d'occurrences dans l'ensemble du corpus, fréquence dans le sous-corpus, fréquence dans le corpus total, nombre de lettres, indice de dispersion et indice de spécificité. Enfin, nous avons ajouté une fonctionnalité pour comparer deux sous-corpus.

3. Premiers résultats

Le premier résultat que nous ayons obtenu est la réalisation du site internet en lui-même, à savoir la création d'une interface dynamique donnant accès aux textes et consentant de les interroger en fonction de différentes variables. En second lieu, plusieurs types de visualisation permettent de représenter l'évolution (dans une perspective globale) ou les rythmes (en fonction des lieux ou du contexte historique et politique de rédaction) des usages langagiers. Les données et leur visualisation permettent de vérifier ou d'écarter certaines thèses que nous avions postulées initialement mais elles ouvrent également la voie à une nouvelle approche des textes qui implique de nuancer certains phénomènes que nous avions identifiés au niveau local. Les tendances de l'écriture qui était invisibles apparaissent alors. L'analyse systématique des relevés d'occurrences et des mesures effectués mettent ainsi en exergue certaines micro-spécificités de la langue telles que le recours constant à la forme hypothétique et aux métaphores de l'incompréhension pour exprimer l'incertitude face aux événements en cours et à venir. L'examen des affinités lexicales a confirmé cette thèse et a permis de mettre au jour une caractéristique déterminante de notre corpus : l'extrême instabilité du monde politique pendant les guerres d'Italie, qui s'exprime à la fois par la centralité du champ sémantique du conflit (qui semble commander les autres champs sémantiques) et par la présence de différents éléments langagiers traduisant une précarité des temps et une fragilisation de la praxis politique.

Bibliography

Blumenthal, P. and Hausmann F. J. (éd). (2006). *Collocations, corpus, dictionnaires. Langue française*, pp. 150.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Scnapp, J. (2012). *Digital humanities*. Massachusetts Institute of Technology.

Genet, J. P. and Zorzi, A. (2011). *Les historiens et l'informatique: un métier à réinventer*. Rome: École française de Rome.

Guillot, C., Heiden, S., Lavrentiev, A., Marchello-Nizia, C. and Rainsford, T. (2013). *La « philologie numérique » : tentative de définition d'un nouvel objet éditorial du point de vue des linguistes*. 27e Congrès international de philologie et de linguistique romanes. https://halshs.archives-ouvertes.fr/halshs-00846767.

Heiden, S. (2004). Interface hypertextuelle à un espace de cooccurrences: implémentation dans Weblex. In Purnelle, G. Fairon, C., Dister, A. (eds), *JADT 2004 - Le poids des mots – Actes des 7ème Journées Internationales d'Analyse Statistique des Données Textuelles*. Presses Universitaires de Louvain, pp. 577-88.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*. n.1 http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1008.

Lasson, M. and Manchio C. (2015). Measuring the Words: Digital Approach to the Official Correspondence of Machiavelli. In Francesca Tomasi, Roberto Rosselli Del Turco, and Anna Maria Tammaro, (eds), *Humanities and Their Methods in the Digital Ecosystem. Proceedings of the Third AIUCD Annual Conference (AIUCD2014). Selected papers*. ACM: New York. http://dl.acm.org/citation.cfm?doid=2802612.2802643.

Lebart, L. and Salem, A. (1994). *Statistique textuelle*. Paris: Dunod. http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html.

Machiavelli, N. Edizione Nazionale delle Opere di Niccolò Machiavelli, *Legazioni. Commissarie. Scritti di governo*. Roma: Salerno: t. I (1498-1500), Marchand, J. J. (eds), 2002; t. II (1501-1503), Fachard, D. et Cutinelli-Rèndina, E. (eds), 2003; t. III (1503-1504), Marchand et Melera-Morettini, M. (eds), 2005; t. IV (1504-1505), Fachard et Cutinelli-Rèndina éd., 2006; t. V (1505-1507), Marchand, Guidi, A. et Melera-Morettini (eds), 2009 ; t. VI (1507-1510), Fachard et Cutinelli-Rèndina (eds), 2011 ; t. VII (1510-1527), Marchand, Guidi et Melera-Morettini (eds), 2012.

Mayaffre, D. (2007). *Philologie et/ou herméneutique numérique: nouveaux concepts pour de nouvelles pratiques?*http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Mayaffre.pdf.

Pincemin, B. (2012). Sémantique interprétative et textométrie. *Corpus*. 10 | 2011. http://corpus.revues.org/2121.

Schöch, C. (2012). Nouvelles configurations : textes, outils, méthodes, et infrastructures de recherche dans les études de lettres. *Configuration(s)*.https://hal.archives-ouvertes.fr/hal-00951518/document.

ThatCamp. (2011). *Manifeste des Digital humanities*. http://tcp.hypotheses.org/318.

Zancarini, J. C. (2007). Une philologie politique. *Laboratoire italien*, pp. 7. http://laboratoireitalien.revues.org/132.

"El Atambor de Plata Suena como Cascaveles de Turquesa". Reconstrucción de la Experiencia Sonora de la Colonización Europea (c. 1480-1650) a Través de un Glosario y un Tesauro Digital

Saúl Martínez Bermejo
saumarti@inst.uc3m.es
Universidad Carlos III de Madrid, Spain

El historiador del sonido se enfrenta a tres retos entrelazados: decodificar la información aural, clasificarla y establecer una sintaxis para dar sentido a la información acumulada (Smith, 2002: 311). Este segundo y tercer retos no deben afrontarse únicamente de un modo personal e impresionista, sino que pueden servir para establecer perspectivas metodologías compartidas, colaborativas y ampliables para la historia del sonido. En esta comunicación corta presentaré un glosario y un tesauro sobre la experiencia sonora en América central durante los siglos XVI y XVII, explicaré la metodología empleada en su creación y discutiré sus distintas aplicaciones.

"El tambor de plata suena como cascabeles de turquesa" es la traducción de un verso náhuatl en el que encontramos un instrumento musical extraño (un tambor hecho de un metal precioso) que se compara con otro artefacto sonoro igualmente complejo, como son los cascabeles de turquesa. El verso muestra tanto la densidad de la información existente en las fuentes, como la dificultad de reconstruir la experiencia sonora a partir de esa documentación. ¿Se trata de instrumentos puramente ficticios, nacidos de la voluntad del poeta, o rituales (pues nada impide que existiesen ejemplares de huethuetl decorados o incluso fabricados con metales preciosos)? Y el verbo frecuentativo *tzitzilica* usado en el original "Xiuhcoyoltzitzilica yn teocuitlahuehuetl" va más allá de un simple sonar y puede traducirse como repicar o repiquetear. El análisis de la poesía náhuatl (León Portilla 2011; Bierhorst, 1985; Damrosch, 1991) puede además ampliarse relacionándolo con otras descripciones de sonidos y paisajes sonoros de la época. Sin renunciar a una reconstrucción narrativa, en el que el verso del tambor de plata se explique en su contexto, un glosario y tesauro digitales ayudan a "clasificar" y "establecer sintaxis" más generales sobre las experiencias sonoras. Si se aprovecha su flexiblilidad para la comparación y el análisis, estas dos herramientas también permiten establecer nuevos modos de comprender y comunicar la información disponible en las fuentes históricas.

Todo glosario es un catálogo de palabras sobre una misma cuestión, con definiciones o comentarios. En su versión digital un glosario no sólo permite una consulta on-line, sino también estructurar la información en diferentes capas o niveles y recuperarla de modo flexible.

Un primer nivel del glosario recogerá las distintas definiciones y traducciones de términos relativos al sonido. Utilizaré para ello diccionarios y gramáticas de época, incluyendo obras bilingües o plurilingües, tales como el *Arte mexicana* del jesuita de origen mexicano Antonio Rincón (1595), el diccionario de Alonso de Covarrubias (1611) o el *Arte de la lengua mexicana* (1645), en el que Horacio Carochi publicó la primera traducción española del verso.

En un segundo nivel, que en realidad es una base de datos, se incluyen ejemplos del uso de estos términos en las fuentes de la época, autor, fecha, localización geográfica, asociaciones entre determinados términos y eventos históricos particulares, valoraciones sobre su significado, su intensidad o su frecuencia, y valoraciones sobre el estado emocional o sensorial que provocan o en el que aparecen. La información recogida en este nivel procede de fuentes como crónicas, las historias naturales, los relatos de viaje, las cartas y los diarios escritos por los participantes en los viajes y campañas militares. De este modo se puede contextualizar la información lexicográfica y se construyen herramientas para desambiguar los términos que aparecen en las fuentes (distinguir, por ejemplo, entre un *huehuetl* y un tambor de dos membranas de origen europeo) y establecer asociaciones entre términos y contextos de uso o entre términos y zonas geográficas.

El tesauro tiene por objetivo reducir las variaciones observadas en el glosario a una serie de categorías lógicas. Para ello se emplearán, naturalmente, términos unificados (campana, por ejemplo) clasificados dentro de una estructura de relaciones ontológicas o semánticas (sonido>artificial>instrumentos musicales>hechos de metal/largo alcance/uso público) A través del uso de estándares como ISO 25964-1 (aún en desarrollo) y de una estructura en XML se favorece la posibilidad de vincular este tesauro a otras clasificaciones previas de efectos sonoros (Chion 1983; Augoyard-Torgue, 1995) o expandirlo a través de colaboraciones de otros investigadores o instituciones.

La construcción del tesauro obliga a reflexionar críticamente sobre el uso de categorías (presentes o de época) para explicar culturas distantes en el tiempo, y se nutre de las perspectivas de la antropología del sonido y de la crítica postcolonial (Tomlinson, 1995; Feld y Brenneis, 2004; Faudree, 2012). Concretamente, plantea hasta qué punto un tesauro permite identificar sobre los cambios en los patrones sonoros o en los grupos de sonidos y sus valoraciones a lo largo de la historia. También pone en cuestión la predominancia de los materiales visuales y textuales y la predominancia de la vista en los modelos teóricos de los historiadores.

El formato digital del glosario y el tesauro ofrece una gran versatilidad a la hora de establecer reconstrucciones paralelas o comparaciones sobre distintas experiencias sonoras. Es posible comparar, por ejemplo, el catálogo de sonidos propio de los niveles perceptivos de un soldado (que no suele aportar muchos datos de carácter musical, pero que ofrece una información relativamente bien contextualizada y localizada en el tiempo y el espacio) con el mundo sonoro de un teólogo o misionero con conocimientos musicales y teóricos sobre el sonido (que categoriza de acuerdo a su propio esquema de división de sonidos y que puede ofrecer variaciones).

La información recogida en el glosario y el tesauro podrá consultarse de varios modos. Junto a una herramienta de búsquedas, una interfaz gráfica permitirá construir "mapas sensoriales" para visualizar la experiencia sonora del pasado. Esta representación novedosa utilizará un diagrama de base circular se colocan varios polos en los que se identifican la procedencia (animal, voz humana, música, artificial, medioambiental), los valores atribuidos a los sonidos (agradable-desagradable, tranquilizador-preocupante, importante-irrelevantes, etc.) y sus contextos (ritual, festivo, actividades diarias, guerra, rebelión, etc.).

Esta investigación es una parte del proyecto *Sound and Silence*, que actualmente desarrollo en la UC3M dentro del programa CONEX-UC3M. El objetivo general del proyecto es investigar los usos y la percepción del sonido durante la colonización portuguesa y española de la edad moderna.

El primer objetivo de *Sound and Silence* es ofrecer una aproximación innovadora y atractiva al estudio interdisciplinar del sonido (Bull & Back, 2003:1-9; Sterne 2012: 3-10; Schmidt 2000: 15-37) . Para ello resaltaré la importancia de la cultura oral (y de las mecánicas de choque e hibridación sonora y musical) en la edad moderna (Greenblatt, 1991: 86-118; Baker, 2008; Irving, 2010) y crearé mapas y/o visualizaciones de la experiencia sonora de un viajero o cronista particular. La página web asociada al proyecto ofrecerá mapas en los que se representarán los tipos de sonidos escuchados durante una expedición, campaña o acontecimiento particular a través de las descripciones existentes en las fuentes históricas creadas por los participantes. Algunos instrumentos, sonidos animales y de artillería y armas de fuego contarán con reproducciones históricas.

Un segundo objetivo del proyecto es investigar las posibilidades contemporáneas de definición semántica de documentación no textual. El texto ha sido y continuará siendo por algún tiempo la base y el centro de la web semántica y de todas las ontologías. En lo que respecta a archivos de música y voz humana la investigación se ha concentrado en los procesos de codificación y reconocimiento. El reconocimiento o la búsqueda por contenidos a partir de huellas sonoras de Shazam o Google se realiza a partir de bases de datos amplias pero limitadas. Sin embargo, el sonido, entendido como un paisaje o un entorno que va más allá la voz y la música plantea retos más amplios, tanto para la teoría historiográfica como para definir la frontera de lo transmisible o "legible" por máquinas.

Hoy día los archivos sonoros de cualquier tipo son accesibles en su gran mayoría a través de un número limitado de metadatos estandarizados referentes a su autor, título, fecha, etc. La identificación de documentos sonoros históricos no es en general mucho más profunda de la que ofrecería una biblioteca como la de iTunes. Herramientas algo más potentes como las folkosomias o la marcación social, permitirían recuperar información sobre el contenido y significado, pero no existe un equivalente histórico para esos modos de categorización. Mi hipótesis es que un glosario que recoja suficientes variaciones contextuales podría permitir de modo similar, comparar los distintos valores atribuidos a diferentes sonidos en determinada época y entender mejor como la gente clasificaba (y clasifica hoy día) el sonido. Un tesauro jerarquizado es, por otra parte, una base para una anotación más profunda de los registros sonoros. La clasificación de archivos sonoros por su contenido se beneficiaria de la existencia de una serie de categorías bien establecidas sobre los tipos de sonido. Ambas herramientas permiten reflexionar sobre los modos en que hoy día organizamos y recuperamos los archivos sonoros y abrir caminos para un tratamiento integral, y diferente, de la información contenida en los formatos de audio.

Bibliography

Augoyard, J-F. and Torgue, H. (eds). (1995). *À l'écoute de l'environnement. Répertoire des effets sonores*. Marseille: Editions Parenthèse.

Baker, G. (2008). *Imposing Harmony: Music and Society in Colonial Cuzco*. Durham, London: Duke University Press.

Bierhorst, J. (1985). *Cantares Mexicanos. Songs of the Aztecs. Translated from the Nahuatl, with an Introduction and Commentary*. Stanford: Stanford University Press.

Bull, M. and Back, L. (eds). (2003). *The Auditory Culture Reader*. Oxford, New York: Berg.

Chion, M. (1983). *Guides des objets sonores. Pierre Schaeffer et la recherche musicale*. Paris: Buchet/Chastel; Institut National de l'Audiovisuel.

Damrosch, D. (1991). The Aesthetics of Conquest: Aztec Poetry Before and After Cortés, *Representations*, **33**: 101-20.

Faudree, P. (2012). Music, Language, and Texts: Sound and Semiotic Ethnography, *Annual Review of Anthropology*, **41**: 519-36.

Feld, S. and Brenneis, D. (2004). Doing Anthropology in Sound, *American Ethnologist*, **31**(4): 461-74.

Greenblatt, S. J. (1991). *Marvelous Possessions. The Wonder of the New World*. Chicago: Chicago University Press.

Irving, D. R. M. (2010). *Colonial Counterpoint: Music in Early Modern Manila*. Oxford, New York: Oxford University Press.

León-Portilla, M. (ed.) (2011). *Cantares mexicanos*, 2 vols. Mexico: UNAM, Fideicomiso Teixidor.

Schmidt, L. E. (2000). *Hearing Things: Religion, Illusion, and the American Enlightenment*. Cambridge (Massachusetts): Harvard University Press.

Smith, B. R. (2002). How Sound is Sound History? A Response to Mark Smith, *The Journal of the Historical Society*, **2**(3-4): 307-15.

Sterne, J. (ed). (2012). *The Sound Studies Reader*. London, New York: Routledge.

Tomlinson, G. (1995). Ideologies of Aztec song, *Journal of the American Musicological Society*, **48**(3): 343-79.

Remediations of Polish Literary Bibliography: Towards a Lossless and Sustainable Retro-Conversion Model for Bibliographical Data

Maciej Maryl
maciej.maryl@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences, Poland

Piotr Wciślik
piotr.wcislik@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences, Poland

Remediation 1.0.: "Printed database"

Polish Literary Bibliography (PBL) is a specialized bibliography which aims to map the totality of literary and cultural life in postwar Poland. It references primarily literary works and literary scholarship, however its entries also cover the related literary critique, adaptations, theatre performances, cinematography, radio and television broadcasts, as well associated events such as conferences or awards. At the heart of PBL lies its subject classification which orders the entries to reflect the domains, hierarchies and entities of Polish literary world, i.e. its ontology in the classical sense. PBL has been developed since 1954 and today covers the period 1944-2001. For most of its history it has existed in print, however since 2000 the data has been collected in the existing digital database which currently covers the period 1988-2001 what gives app. 600 000 records.

The vicissitudes of PBL remediations could be accurately captured through an urban planning metaphor. What definitely strikes every visitor to a large Moroccan city is a great contrast between *medina*, the traditional old town with centuries-long history, and *Ville Nouvelle*, new district built under the French Protectorate in the first half of the 20th century. The former reminds a maze with endless narrow streets, and buildings which are stuck densely next to each other with no visible order, whereas the latter is the essence of modern architecture with wide boulevards, large buildings and streets laid out in a grid pattern.

The current online database is quite exemplary for early bibliographical and cataloguing projects (in Poland as elsewhere) in that it is geared towards remediating the print form of the PBL instead of taking advantage of the new medium (cf. Antelman, Lynema and Pace 2006, 128). It is a tailor-made relational database developed in Oracle whose data model is built on a plethora of dataspaces for different types of records, accompanied by various catalogues of creators, contributors, associated institutions and subject headings. The former set reflects PBL's main entities: literary works in monographs and journals, adaptations in cinematography, radio and television and associated events. The latter represents an early digital take on the index card catalog, the traditional tool of the bibliographer. Furthermore each record has a special markup in order to assure that its display at the frontend follows the structure of the paper edition.

The result of this remediation is a *medina*-like database, very rich and complicated but not fit for modern uses. It makes perfect sense for people who built it, yet at the same time it is difficult to navigate by those lacking the local knowledge - be it a human or the machine. As it often happens with relational databases,[1] it does not comply with any of the common standards in terms of record structure or data formats, what eventually leads to serious problems with both preservation and interoperability of collected data.

Towards remediation 2.0

The aim of the research project we are currently pursuing (*Polish Literary Bibliography – a knowledge lab on contemporary Polish culture*) is to reestablish the PBL database project on Linked Open Data principles for its better reuse within and beyond the bibliographic domain (see e.g. Roszkowski 2013; Coyle 2010). However, we want to do better than the French colonizers of Morocco. The modernisation of PBL will be reflexive insofar as it will reconcile the OWL and the PBL's unique ontology of the literary world expressed through the structure of its entries and metadata. The main task of the current phase of the project is development and application of the new data model. This task involves (1) the choice of vocabularies and ontologies and (2) rendering of the subject classification structure.

(1) Vocabularies and ontologies (in the narrow sense used in information science) are needed to disambiguate the RDF triples (subject-predicate-object expressions). Here we need to balance two criteria. First the vocabularies and ontologies must enable widest possible sharing in the data cloud. Second they must be granular and complex enough in order to reflect the PBL data model, since adding too many heterogeneous elements would be counterproductive. The above applies to both metadata elements and their values.

Whereas the choice of value vocabularies was rather straightforward, using the geonames and Virtual International Authority File (VIAF) for disambiguating geographical, personal and corporate names, the choice of the meta-ontology,[2] or the vocabulary describing the metadata elements of the current PBL data model was much more difficult. It would be only natural to opt for one of the ontologies dedicated for describing bibliographic records, such as Functional Requirements for Bibliographic Records (FRBR) and its Resource Description and Access (RDA) and Bibliographic Framework Initiative (BIBFRAME) vocabulary variants (cf. Coyle 2016). Indeed, both contain a crucial distinction between "works" (a certain intellectual creation as such, regardless its edition, format or medium) and "instances" (expressions and manifestations of this intellectual creation) which in PBL is paramount for referencing editions, adaptations and critiques of a literary oeuvre of a particular author. For example, a review of *Don Quijote* refer to either Cervantes' literary achievement in general or to the newest translation of the Spanish original into Polish.

However, the FRBR-based ontologies are either not well equipped to handle theatre, cinematographic, radio and television instances of literary works, or (as in the case of FRBRoo) too complex to be easily handled by metadata producers in their everyday practice (Coyle 2016, 153).[3] Therefore, we opted for a solution that is more generic but robust enough - the schema.org ontology. However contestable due to its rather restricted vocabulary when it comes to describing books, this solution is not unprecedented in the bibliographic domain.[4]

This process of mapping is by no means mechanical. In many cases the PBL original methodology and the solutions of the first remediation entailed conceptual challenges, which will be addressed in more detail in our presentation. For instance, one needs to solve the tension between a minute bibliographic description on one hand, and the standard vocabulary on the other. Expressions entailing similar yet slightly different properties of the book such as "woodcut engravings"; "illustrations"; "drawings"; "reproductions"; "pictures"; "prints" need to be fit into the elements of the formal vocabulary of schema.org, properties such as "illustrator" and "artform."

(2) The second challenge of the new data-model involves the PBL subject classification structure. Here the option of using one of the existing and well-established subject headings/authority files published as Linked Data, such as the Library of Congress or German National Library Subject Headings was rather out of question, given the methodological uniqueness of PBL. Instead we will strive to create our own Linked-Data ready classification scheme while at the same time providing a partial mapping to existing resources.

To realize the scope of this challenge one needs to bear in mind that PBL has been an ongoing project for the last sixty years. During this time, not only literary life and its study have evolved (cf. the emergence of the online literary life, Maryl 2015), but also certain state entities disappeared (e.g. Yugoslavia or the Soviet Union). Given that the future database will be populated through retroconversion of the paper records in addition to the existing database records, we cannot take the current classification for granted, but also accommodate its historical evolution. A non-intrusive way to account for the historicity of PBL would be to add timestamps to subject headings. Whether a synthetic data-reconciliation layer is possible requires further analysis.

Conclusions

In the concluding remarks we will concentrate on the expected benefits of translating PBL into LOD.

- PBL datasets can be enriched through integrating other Linked Data collections (e.g. geographical data on places relevant to literary life).
- Data exchange protocols can be established between PBL and other bibliographies published as Linked Data.
- PBL data can be used for data-driven research in the humanities on such fields as reception history or transfer studies.
- The methodology and the production pipeline developed in this project can be reused for retroconversion of other disciplinary bibliographies.

Acknowledgment

This work was supported by Polish Ministry of Science and Higher Education through the National Programme for the Development of the Humanities (grant number: NR 0061/NPRH3/H11/82/2014).

Bibliography

Antelman, E., Lynema, A. and Pace, K. (2006). Toward a Twenty-First Century Library Catalog *Information Technology & Libraries*, **25**(3): 128–39.

Coyle, K. (2010). Understanding the Semantic Web: Bibliographic Data and Metadata, *Library Technology Reports*, **1**: 5-31.

Coyle, K. (2016). *FRBR Before and After*. Chicago: ALA Editions.

van Hooland, S. and Verborgh, R. (2014). *Linked Data for*

Libraries, Archives and Museums: How to clean, link and publish your metadata. London: Facet Publishing.

Maryl, M. (2015). *Życie literackie w sieci: pisarze, instytucje i odbiorcy wobec przemian technologicznych*, (Literary life online: writers, institutions and readers facing technological changes). Warszawa: Wydawnictwo IBL.

Roszkowski, M. (2013). Od MARC 21 do Semantic Web: reprezentacja metadanych bibliograficznych w środowisku sieciowym (From MARC21 to Semantic Web: bibliographic metadata representation in network environment). In Franke, J. (ed.), *Bibliografi@: źródła, standardy, zasoby*. Warszawa: Wydawnictwo SPB, pp. 13-37.

Stahmer, C. (2015). Approaches to Digital Bibliography and Book History: Ontology Reference, *Carl Stahmer blog*. http://www.carlstahmer.com/2015/06/approaches-to-digital-bibliography-and-book-history-ontology-reference/ (accessed 29 February 2016).

Notes

[1] For a useful comparison of different approaches to data modelling, see van Hooland and Verborgh 2014: 11-70.

[2] Our understanding of "meta-ontologies" (as opposed to value vocabularies) is analogous to what the W3C Library Linked Data Incubator Group calls "Metadata Element Sets". An additional feature of a meta-ontology is that it " combines and organizes other ontologies to describe objects in a complex way." See: Stahmer 2015.

[3] In case FRBRoo is in fact widely adopted by the community of practice, we will be definitely interested in mapping it to our data model.

[4] For list of library-specific extensions proposed by the Schema Bib Extend W3C Community Group, see bib.schema.org. WorldCat.org uses a subset of schema.org terms for Linked Data (https://www.oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html). Updates on schema.org developments are frequently announced on Richard Wallis' blog dataliberate.com

The Digital Scholarship Training Programme at British Library: Concluding Report & Future Developments

Nora McGregor
nora.mcgregor@bl.uk
British Library, United Kingdom

Mia Ridge
mia.ridge@bl.uk
British Library, United Kingdom

Stella Wisdom
stella.wisdom@bl.uk
British Library, United Kingdom

Aquiles Alencar-Brayner
Aquiles.Alencar-Brayner@bl.uk
British Library, United Kingdom

Digital Scholarship Training Programme 2012–2015

Research libraries and cultural heritage institutions must be able to adapt to a changing research landscape and invest in the development of staff skills and core competencies to match if they are to continue to effectively support and engage with modern scholars (Adams, 2013). The Digital Research team at British Library, which includes the BL Labs project and the Digital Curator team, engages with those operating at the intersection of academic research, cultural heritage and technology to enable innovative use of our digital collections, and creates opportunities for library staff to develop skills necessary to support emerging areas of modern scholarship, particularly the Digital Humanities (DH).

This paper presents the final report on the pilot Digital Scholarship Training Programme delivered to staff at British Library between 2012-2015. It provides an evaluation of the skill-building initiative, incorporating preliminary findings from research into the current trends and international developments in the field of DH that will inform the next phase of staff training.

In 2012, the Digital Curator team embarked on a plan to design and deliver a bespoke training programme for staff (McGregor et. al, 2013). Four objectives were set to guide and ultimately measure the success of the programme:

- Staff across all collection areas are familiar and conversant with the foundational concepts, methods and tools of digital scholarship.
- Staff are empowered to innovate.

- Collaborative digital initiatives flourish across subject areas within the Library as well as externally.
- Our internal capacity for training and skill-sharing in digital scholarship are a shared responsibility across the Library.

In consultation with experts from within the Library and institutions on the leading edge of digital scholarship we designed and delivered in-house a catalogue of 19 one day courses suited to building the digital skills of information professionals in the research library and cultural heritage sector. Though much of the programme was rooted in and inspired by the field of digital humanities, the wider umbrella of Digital Scholarship was retained to future-proof the programme and to envelope developments in related fields like social and computer sciences. The following titles represent a cross-section of courses created:

- Behind the Screen: The Basics of the Web
- Crowdsourcing in Cultural Heritage (Ridge, 2015)
- Georeferencing and Digital Mapping
- Social Media: An Introduction to Blogging and Twitter
- Working with Digital Objects: From Images to A/V
- Information Integration: Mash-ups, APIs and the Semantic Web (Stephens, 2014a)
- Cleaning up Data (Stephens, 2014b)

The four-member Digital Curator team oversaw the running of 88 hands-on courses (or roughly 30 a year) between November 2012 and September 2015. Courses were delivered by a mix of internal and external trainers. Over 400 individual staff members came through the programme, on average attending two or more courses each.

Throughout the pilot we collected feedback formally via post-course evaluation forms, and informally through avenues such as a weekly Digital Research Clinic, and personal conversations that arose in the course of our daily work. Colleagues were asked to provide comments to help us improve the course, including what they enjoyed most out of the day, what they anticipated using in their work, and what was not clearly articulated.

Hands-on practical exercises were cited most often as the most enjoyable element, though not to the exclusion of the lecture and discussion time which participants felt provided necessary context. While many attendees cited specific technologies such as Open Refine as something they anticipated using in their work, they also tended to comment that having the technology underpinning innovative digital research projects demystified was helpful inspiration for future projects. Topics which could have been more clearly articulated centred on a lack of clarity on practical steps for turning aspiration into application.

Looking specifically at the objectives set, there is ample evidence to support the continuation of the training programme, such as the incorporation of the programme in staff induction for established Library projects such as Qatar Digital Library. As staff across all collection areas have become more familiar and conversant with the foundational concepts, methods and tools of digital scholarship, we have witnessed its profile increase across the Library. For instance three new PhD placements were offered this year specifically within the digital research domain (Sheperd, 2016) for the first time.

Staff have felt empowered to innovate, and collaborative digital initiatives have flourished. A particularly cogent example is that of curator Dr. Sandra Tuppen, who attended one of our courses on cleaning up data and went on to secure a £79,000 grant towards a research project which enriched and cleaned British Library catalogue data so that it could be successfully aligned with other printed music datasets in support of a big data approach to the history of music (Tuppen, 2014).

Our internal capacity for training and skill-sharing in digital scholarship has become a shared responsibility across the Library, with internal course instructors now outnumbering external instructors. With the increasing number of Library staff being involved in projects and other programmes that include digital research activities and methodologies, we have been able to integrate more in-house expertise to the courses offered. Additionally the Digital Curator team has prioritised our own upskilling through a monthly informal “Hack & Yack” where we work through online tutorials with a view towards incorporating them into training, as well as more formal courses such as Train the Trainer aimed at enhancing our teaching strategies, combining theory and practical methodologies in the planning and delivery of the courses.

Looking to the future

We will continue the Digital Scholarship Training Programme and for 2016/2017 will maintain 7 of the 19 courses in their current form:

1. 101 This is Digital Scholarship
2. 103 Digitisation at the British Library
3. 105 Crowdsourcing in Libraries, Museums and Cultural Heritage Institutions
4. 108 Geo-referencing and Digital Mapping
5. 109 Information Integration: Mash-ups, API’s and Linked Data
6. 116 Metadata for Electronic Resources
7. 118 Cleaning up Data

Focusing on delivering this smaller core of courses will free up resource to improve upon how we:

- Reach staff who are keen and could most make use of the information but have not yet engaged
- Providing guidance and support to staff who are looking to implement what they have learned
- Addressing more explicitly the challenges and opportunities for working with complex collection materials, such as with non-Western materials

We aim to provide a more diverse training offering to ensure that there are sufficient opportunities for staff in a variety of roles at the Library to engage with digital research. Often colleagues have said they would like to attend a course, but either their workload is such that they feel they cannot spare a full day for it or they work a rota, as is the case with our colleagues who staff reading rooms. Finding creative ways to articulate more clearly and succinctly the practical value of time spent on a course, for example through shorter more frequent taster sessions explaining how a digital tool or method might help solve a specific problem, may help to reach those who have yet to engage. We are also working in partnership with reading room staff on rota to develop new approaches for conveying the training (perhaps through short informational videos).

On the opposite end of the spectrum is the need to support increasing numbers of library staff who have engaged with the programme and are now looking to implement what they have learned. In an ideal world we could offer 'just in time' training to colleagues at the point of their immediate need. However, as Miriam Posner (2012) and others have discussed, each question a colleague asks may bring with it a hidden overhead of time taken to respond well. In some cases we may seek to hire existing trainers to support specific project needs but more practically, we will look to better promote and leverage our weekly Digital Research Clinic, a drop-in service for staff to get guidance on any aspect of digital research. A collection of practical 'Getting Started' guides will be further developed and shared via an internal wiki. Additionally, Digital Curators sit on key infrastructure development projects so as to directly inform the development of these in support of digital research.

Finally, digital scholarship is a complex and global affair, as evidenced by the rapid expansion of DH centres around the world. Our courses to date have tended to deal with relatively simple forms of digitised material such as digitised printed English language books. However for colleagues working with non-Western texts for instance, knowledge of cutting edge developments in transcription and Optical Character Recognition would be highly beneficial in helping ensure these materials can be leveraged by digital scholars. Our collections are as global and diverse as the DH communities interests worldwide, and future staff training provision must more accurately address the complexities and opportunities of working with our vast non-Western materials.

Bibliography

McGregor, N. and Farquhar, A. (2013). The Digital Scholarship Training Programme at British Library, *Abstracts | Digital Humanities 2013*. http://dh2013.unl.edu/abstracts/ab-264.html (accessed 16 February 2016).

Posner, M. (2012). What are some challenges to doing DH in the library? *Miriam Posner's Blog*. http://miriamposner.com/blog/what-are-some-challenges-to-doing-dh-in-the-library/ (accessed 29 February 2016).

Ridge, M. (2015). Resources for "Crowdsourcing in Libraries, Museums and Cultural Heritage Institutions", *Mia Ridge's Blog*. http://www.miaridge.com/resources-for-crowdsourcing-in-libraries-museums-and-cultural-heritage-institutions/ (accessed 16 February 2016).

Shepard, J. (2016). PhD placements in Digital Scholarship British Library, *Digital Scholarship*. http://britishlibrary.typepad.co.uk/digital-scholarship/2016/02/phd-placements-in-digital-scholarship.html (accessed 16 February 2016).

Stephens, O. (2014). Information Integration: Mash-ups, APIs and the Semantic Web | Overdue Ideas, *Information Integration: Mash-Ups, APIs and the Semantic Web*. http://www.meanboyfriend.com/overdue_ideas/2014/10/information-integration-mash-ups-apis-and-the-semantic-web/ (accessed 16 February 2016).

Stephens, O. (2014). Working with Data using OpenRefine | Overdue Ideas, *Working with Data Using OpenRefine*. http://www.meanboyfriend.com/overdue_ideas/2014/11/working-with-data-using-openrefine/ (accessed 16 February 2016).

Tuppen, S. (2014). A Big Data History of Music, *British Library Music Blog*. http://britishlibrary.typepad.co.uk/music/2014/04/a-big-data-history-of-music.html (accessed 3 March 2016).

Building Large Persons' Networks to Explore Digital Corpora

Giovanni Moretti
moretti@fbk.eu
Fondazione Bruno Kessler, Italy

Sara Tonelli
satonelli@fbk.eu
Fondazione Bruno Kessler, Italy

Stefano Menini
menini@fbk.eu
Fondazione Bruno Kessler, Italy; University of Trento, Italy

Introduction

Although representing large corpora through the network of persons' interactions has become quite popular in the Digital Humanities community (Elson et al., 2010), several parameters can have an impact on the resulting network, especially when it is automatically extracted. In this work, we present a step-by-step procedure to extract persons' networks from documents and select possible configurations in order to increase readability and ease

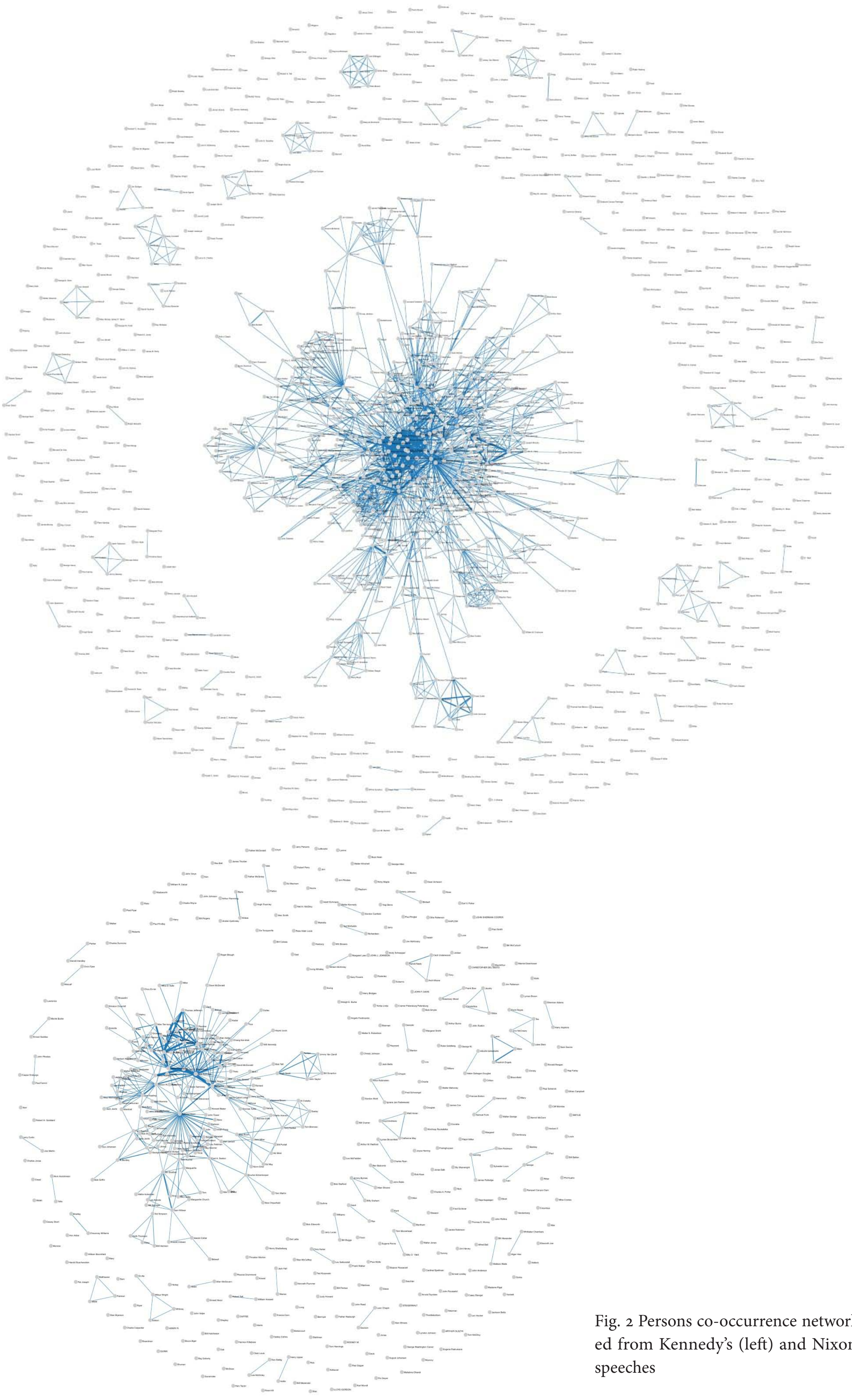


Fig. 2 Persons co-occurrence network extracted from Kennedy's (left) and Nixon's (right) speeches

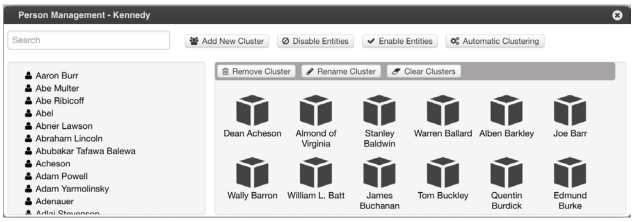
the interpretation of the obtained information. We also discuss some open issues of the task.

We rely on the same assumption as for word co-occurrence networks: *two persons who tend to be mentioned together in a corpus share some commonality or relation from the author's perspective.*

The Methodology

We implemented a novel tool for the automated extraction of a persons' network from a corpus in the online ALCIDE platform[1] (Moretti et al., 2014). The module is based on the following steps: *i)* the corpus is first analysed with the Stanford named entity recognizer (Finkel et al., 2005), in order to recognize persons' mentions (e.g. *John Kennedy*, *F.D. >Roosevelt*, etc.). In the network representation, we assume that persons correspond to nodes and edges express co-occurrence within a given token window; *ii)* We build a person-person matrix where we assign an edge weight of 1 every time two persons are mentioned together within a certain context window. Every time a co-occurrence is repeated, the weight is increased by 1; *iii)* The final output is a weighted undirected network where edge weights are the co-occurrence frequency. In the default configuration, name mentions are collapsed onto the same network node only if they have an exact match. In order to allow a more flexible creation of the network, a "Person Management" functionality (Fig. 1) has been implemented, through which users can collapse nodes referring to the same person (e.g. *J.F. Kennedy* and *John Kennedy*). This manual check is done through an interface, without the need to access directly the underlying matrix.

From a technical point of view, the information is stored in a relational database management system, in order to grant multi-user access, good performances and high flexibility regarding the queries. The persons' co-occurrence matrix is visualized as a network by means of the d3 javascript framework[2]. During the conversion of the matrix in the json used by d3, the nodes are enriched with additional information, such as the list of documents containing the corresponding entity and the number of connections.



Aaron Burr
Abe Multer
Abe Ribicoff
Abel
Abner Lawson
Abraham Lincoln
Abubakar Tafawa Balewa
Acheson
Adam Powell
Adam Yarmolinsky
Adenauer
Adlai Stevenson

Dean Acheson, Almond of Virginia, Stanley Baldwin, Warren Ballard, Alben Barkley, Joe Barr
Wally Barron, William L. Batt, James Buchanan, Tom Buckley, Quentin Burdick, Edmund Burke

Fig. 1 View of the Person Management tool

Some settings such as the co-occurrence window type (sentence or token) and width (number of sentences/tokens) are arbitrary, although they have a relevant impact on the extracted network and on its readability. Therefore, the system gives the possibility to change such settings and regenerate the co-occurrence matrix at runtime. In the following sections we will discuss some of these parameters and explain their impact in the light of a use case related to Nixon and Kennedy's speeches of the 1960 presidential campaign. The corpus consists of 282 documents by Nixon (830,000 tokens) and 598 documents by Kennedy (815,000 tokens)[3]. All networks displayed in the following sections are screenshots of the system output and are dynamically displayed.

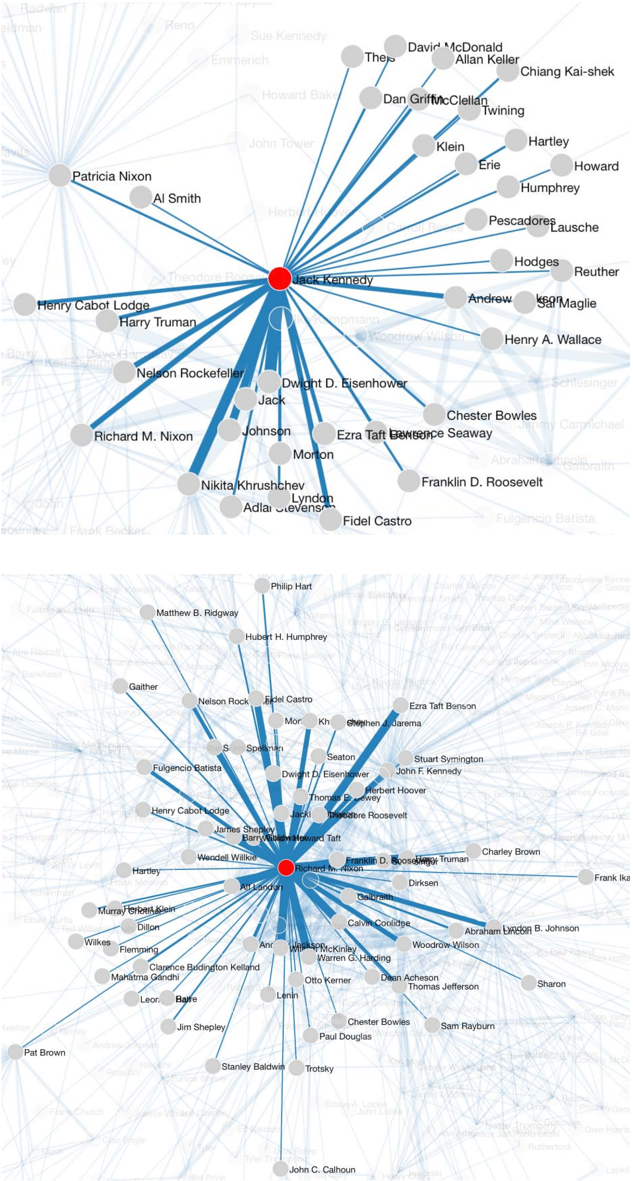


Fig. 3 Co-occurrence network of "Nixon" mentions extracted from Kennedy's corpus (left) and of "Kennedy" mentions from Nixon's (right) speeches

Default configuration

In our default configuration, the tool extracts persons' networks using 1 sentence as a co-occurrence window and collapsing on the same node only name mentions with an exact match. As shown in Fig.2, this basic configuration is enough to highlight the differences between Kennedy's and Nixon's networks: the first is much larger and much more connected, with several cliques that tend to emerge from the overall picture. Nixon's network, instead, is smaller (i.e. less persons are mentioned in his corpus) and less dense.

By zooming in the pictures, it is possible to focus on single nodes of interest. For example, if we compare Nixon's mentions appearing in Kennedy's speeches, and the other way round (Fig. 3, left and right resp.), we observe that in both cases the opponent is frequently mentioned with 'enemies' of the time such as Fidel Castro and Khrushchev. However, this association with negative figures is much more frequent in Kennedy's speeches (e.g. Nixon is mentioned also with Trotsky and Lenin), probably because Nixon had already a prominent role in US foreign policy being Vice-President.

Changing configuration parameters

The tool allows users to move from the default configuration to a more customizable one, where it is possible to change the type (sentence or token) and the size of the context window taken into account for the co-occurrences as well as set a threshold to the edges' weight (number of co-occurrences). By tuning these parameters, it is possible to transform the networks presented in Fig. 2 to obtain a more readable representation, filtering minor nodes and emphasizing information previously hidden by the large amount of information.

Reducing the co-occurrence window increases the probability to extract persons that are more strictly related. At the same time, by increasing the minimum edge weight threshold, we reduce the information visualized, filtering out all the persons co-occurring only once in favour of persons co-occurring consistently through the entire corpus.

Fig. 4 shows the networks, obtained from the data in Fig. 2, generated by setting the maximum token range to 10 (thus, on average, less than the sentence length adopted in Fig. 2) and the minimum edge threshold to 2. The result is a visualization with less but more readable data. In Nixon's network, we can easily spot some well-defined clusters such as the one grouping the leaders of the communist world (i.e Khrushchev, Stalin, Mao Tse-tung), the cluster of the main representatives of international politics in 1960 (e.g. de Gaulle, Nehru, Adenauer) or a cluster reflecting Nixon's attitude to refer to previous U.S. Presidents (e.g Andrew Jackson, Thomas Jefferson). Also the network from Kennedy's corpus is more understandable, including a cluster with prominent Communist politicians (e.g. Khrushchev, Castro, Kadar), but also clusters defining local democratic representatives, for instance those from California (e.g. Pat Brown, John Moss) or those from Pennsylvania (e.g. David Leo Lawrence, Joseph S. Clark).

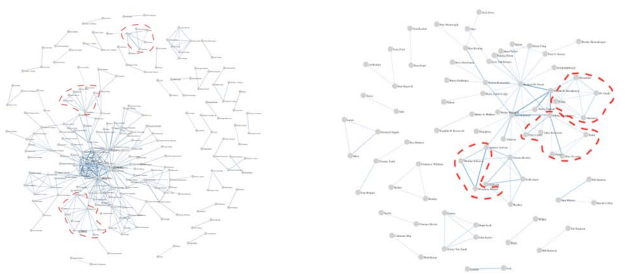


Fig. 4 Persons co-occurrence network extracted from Kennedy's (left) and Nixon's (right) speeches using a 10 tokens windows and a minimum threshold of 2.

Discussion

As shown in the above examples, building a persons' network in an automated fashion implies making some a-priori choices that strongly affect the outcome of the analysis. Such choices are influenced by the type of analysis required by the user. If *distant reading* is the main goal, the parameters proposed as default by our tool, as shown in Fig. 2, seem to be appropriate. This analysis gives an overview of the overall network dimension and density, and makes it possible to compare two networks at a glance. Instead, if *close reading* and a qualitative analysis of the connections are more relevant, reducing the window width and displaying only the most connected nodes is necessary. Since in a typical research scenario distant and close reading are both present, and users need to zoom in and out frequently, a tool that changes the network on demand in real time should be implemented. In this respect, Gephi (Bastian et al., 2009), probably the most widely used tool for network analysis in the Digital Humanities community, shows some limitations. Although it provides useful in-built metrics for analysing a network structure, it does not offer the possibility to test different parameters on the fly. Also its integration in an online, browser-based environment is quite complex, as well as the connection to a text analysis pipeline.

Other issues related to the automated creation of persons' networks are worth mentioning. Since natural language processing tools are involved in the pre-processing step, users should be aware of the possible mistakes introduced by this pipeline. In particular, Named Entity Recognizers may label as persons other types of entities, wrongly introducing nodes in the network. Other possible mistakes are more difficult to spot and concern homonyms (i.e. nodes that should be split). Such cases can be solved only resorting to a cross-document entity coreference system, where mentions can be resolved by linking them to the entities they refer to. A finer-grained outcome could be achieved integrating also an intra-document coreference system, able to link mentions referring to the same person in a text. Cross- and intra-document coreference would be necessary to ensure that all persons mentioned in a corpus are included in the extracted network. Nevertheless, the impact of automated processing on the quality of the network needs to be further investigated.

Bibliography

Bastian M., Heymann S. and Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media (ICWSM).*

Elson, D. K., Dames, N. and McKeown, K. R. (2010). Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden.

Finkel J.-R., Grenager T. and Manning, Ch. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-70.

Moretti G., Tonelli S., Menini S. and Sprugnoli R. (2014). ALCIDE: An online platform for the Analysis of Language and Content in a Digital Environment. In *Proceedings of the First Italian Conference on Computational Linguistics (CLIC-2014)*, Pisa, Italy.

Notes

[1] http://celct.fbk.eu:8080/Alcide_Demo/
[2] http://d3js.org
[3]http://www.presidency.ucsb.edu/1960_election.php

Collaborative Translation Using FromThePage

Laura Morreale
lmorreale3@fordham.edu
Fordham University, United States of America

Barbara Mundy
mundy@fordham.edu
Fordham University, United States of America

Thomas O'Donnell
todonnell12@fordham.edu
Fordham University, United States of America

Nicholas Paul
npaul@fordham.edu
Fordham University, United States of America

Brian Reilly
breilly17@fordham.edu
Fordham University, United States of America

Ben Brumfield
benwbrum@gmail.com
FromThePage

Many pre-print texts emerge from an authorial context that was collective rather than individual. Pre-print texts may be the result of an accretive process, where portions of the work were added or subtracted to conform to the needs of copyist or audience, and where the surviving versions can reflect the contribution of many different minds (Fisher, 2012). A truly sensitive translator of a text from the past must take into account the historical setting of the written product, what has been called the *social logic of the text* (Spiegel, 1990) to interpret the words for the reader and convey the meaning for those who wrote them and for their contemporary audiences (Paul, 2008).

Just so, the job of the translator has become increasingly complex as our notion of text also expands. Ideally, the process of translating pre-modern texts would mirror their progressive creation, incorporating multiple voices and layers of interpretation. With the integration of digital modes of thinking and acting into our daily lives, what some digital theorists have called "eversion" (Jones, 2014), new textual communities in some way mimic those that participated in the collective process of pre-print text creation. Digital tools have made it possible to assemble voices in a virtual community of textual reception, one in which participants need not be in the same physical space to exchange ideas in real time about the meaning of a text.

Scholars at Fordham University have embarked upon two different translation projects that are sensitive to elements of pre-print textual creation and that aim to forefront the process in ways understood by digitally-inflected communities. The projects emerge from a pragmatic vision of work-sharing, yet simultaneously co-opt our increasing comfort with virtual communities to recreate the collective context of pre-print texts. These projects come from unrelated disciplines and cultural contexts: the first brings together scholars with linguistic, literary, and art historical expertise to produce a new translation of the Codex Aubin, a manuscript written in Nahuatl created between 1576 and 1609 by indigenous scribes in Mexico City; and the second, led by the French of Outremer Legal Texts Working Group, brings together nine scholars to examine and translate three thirteenth-century legal texts written in a type of Old French from the Holy Land.

Although digital tool-builders have recognized the power of the collective translation process from the standpoint of a division of labor, most translation platforms are designed to manage work-flow, not to accommodate incertitude, discussion, or to express a plurality of interpretations (Gambier, 2014). Using the norms of crowdsourcing as a point of departure, Fordham scholars have worked with program-developer Ben Brumfield to substantially extend his open-source digital edition tool FromThePage (http://www.fromthepage.org/) to allow groups of scholars to translate selected texts in a communal fashion. For example, since the 1841 editions of the French-language legal texts were digitized by the Internet Archive, the existing FromThePage-Internet Archive integration was extended to ingest the OCR produced by the digitization process as a starting point for page transcripts (Beugnot, 1841). The translation process itself required modifying the data model to support parallel texts for each transcribed page, as well as new user interfaces for editors to toggle between transcript and facsimile while translating. Since users may be unfamiliar with Old French or Nahuatl, the presentation interface was revised to support a translation-first experience.

Translation teams working on both Fordham FromThePage projects include language specialists, historians, and content experts who bring their proficiencies to the discussion, but who will nonetheless disagree along the way. The assembled communities are closed and highly selective; our end product will be a curated conversation that reflects the specialization of each team member. Using FromThePage gives us the option to make the deliberations transparent to our end-users in a way that remains epistemologically true by mapping the process of consensus necessary for a final translation. Reaching an agreement on our final versions, for example, may be the result of one scholar arguing for a reading according to his or her own expertise, or of a collective ignorance and the need to choose the most neutral term possible; our final translations will expose this decision-making process and situate our users in the midst of communal textual production.

Achieving this transparency required a number of unanticipated changes. The first concerned the commenting mechanism: while editors transcribing a manuscript often need to discuss unclear handwriting, translators' discussions of variant readings need to be surfaced for readers. In addition, parallel texts required a total re-write of the full-text search mechanism to handle verbatim text of the original, editorial emendations, and English translations. Finally, the nature of the sources revealed fundamental challenges in the representation of text within FromThePage; since the page division of the *Assises* is an artifact of 19th-century typesetting, semantic divisions had to be de-coupled from the facsimile pages for online readers and consumers of the exported TEI-XML documents. All of these enhancements were re-incorporated into the FromThePage source code, allowing other digital edition projects to benefit from the new translation and OCR correction features. The projects underscore the communally-crafted nature of our chosen texts by marking conflicting opinions, and the results will serve as a test-case for other kinds of collaborative textual projects, particularly those that contain non-standard languages or terminologies.

Bibliography

Legal Texts project: http://fromthepage.ace.fordham.edu/collection/show?collection_id=1

Codex Aubin project: http://fromthepage.ace.fordham.edu/collection/show?collection_id=2

Beugnot, A. (ed). (1841). *Assises de Jérusalem, ou, Recueil des ouvrages de jurisprudence, composées pendant le xiiie siecle dans les royaumes de Jérusalem et de Chypre*. Recueil des historiens des Croisades. Lois, Paris: Imprimerie royale, **1-2**.

Fisher, M. (2012). *Scribal Authorship and the Writing of History in Medieval England.* Ohio State University Press.

Gambier, Y. (2014). "Changing Landscape in Translation," *International Journal of Society, Culture and Language*. http://www.ijscl.net/article_4638_848.html. **2**(2): 1-12.

Paul, G. (ed). (2008). *Translation in Practice: A Symposium by the British Centre for Literary Translation.* Dalkey Archive Press.http://www.llvs.lt/img/File/Translation_in_Practice_book.pdf.

Jones, S. E. (2014). *The Emergence of the Digital Humanities.* Routledge.

Spiegel, G. (1990). "History, Historicism, and the Social Logic of the Text in the Middle Ages," *Speculum*, **65**(1): 59-86.

Linguistic Variation In The Hebrew Bible: Digging Deeper Than The Word Level

Martijn Naaijer
m.naaijer@vu.nl
Vrije Universiteit Amsterdam, Netherlands, The

Dirk Roorda
dirk.roorda@dans.knaw.nl
Data Archiving and Networked Services

One of the most discussed issues in Biblical studies in the past 15 years is the history of Biblical Hebrew. Standard works on this issue assume that one can distinguish between Archaic Biblical Hebrew, Early Biblical Hebrew and Late Biblical Hebrew (Saenz-Badillos; 2004, Hurvitz; 2013). This position has been challenged in a number of recent publications, in which the authors state that the variation that can be found in Biblical Hebrew is better explained by assuming that there have been different styles of Biblical Hebrew (Young, Rezetko and Ehrensvärd, 2008) in use throughout the biblical period, which is roughly the whole first millennium BCE. A complicating factor in the research on the biblical texts is that we know relatively little about their transmission in the centuries after their composition.

In general the manuscript used for research on Biblical Hebrew is the Codex Leningradensis, which was created in 1008/1009 CE. There exist older manuscripts of the Hebrew Bible; by far the most important ones are those found in the Qumran Caves which can be dated to the beginning of the Common Era, but many of these manuscripts survive in a fragmentary state.

Many studies on the diachrony of Biblical Hebrew are concerned with Hebrew vocabulary. These have resulted in long lists of early lexical items that were supposed to be replaced gradually by late alternatives. These late alternatives can often be identified as loans from languages like Aramaic and Persian (Young, Rezetko and Ehrensvärd, 2008).

One of the problems of studying vocabulary as a gauge of linguistic change is that the vocabulary could have been manipulated easily during the process of transmission. Scribes could change words, thereby consciously archaizing the language of a text.

In order to solve this problem, we study syntax instead of vocabulary. Forming sentences takes place on a less conscious level than choosing words, and therefore this is a better way of studying continuity and change in the history of Biblical Hebrew. In our study we investigate the use of prepositions accompanying a whole range of verbs. In the literature on linguistic variation in Biblical Hebrew the use of prepositions and other function words in various contexts has been studied before, but this has always been done in a very restricted way. Sometimes only a few biblical texts had been studied or the data had been extracted from one manuscript exclusively (Hornkohl, 2014:218-38).

In our research we will focus on verbs of motion and on stationary verbs. In the former category we find verbs like אוב (bōʔ, to come), הלע (ʕālā, to go up) and אצי (yāṣā, to go out), in the latter we find verbs like בשי (yāšav, to sit) and דמע (ʕāmaḏ, to stand). These verbs have in common that in most cases they have a locative as complement, which is often introduced by a preposition (Oosting, Dyk and Glanz, to be published). It is known that various prepositions can be used with a given verb and this variation can be found in parallel texts in the Codex Leningradensis, within specific biblical books and between different manuscripts (Kutscher, 1974).

The use of function words like prepositions is well known in authorship attribution (Argamon and Levitan; 2005, Garcia and Martin; 2007, Segarra, Eisen and Ribeiro; 2013), but in the case of ancient religious texts, detecting the author of a text is a controversial issue. Not only could texts have been adapted during the transmission of the complete text, also the composition of a text may have had a long history. Therefore we do not try to find the supposed author of a text, but based on the study of prepositions accompanying verbs of motion we would like to find out what is the main factor of the variation in the use of these prepositions. Is it related to diachronic development of the Hebrew language or to the way the texts were transmitted or both or are there still other options? We investigate these issues by comparing the thousands of instances of prepositions accompanying verbs of motion and stationary verbs in:

1. different books in the Codex Leningradensis (Genesis, Exodus, etc.)

2. different manuscripts (e.g. compare Isaiah in the Codex Leningradensis with the Great Isaiah Scroll)

3. parallel texts.

This kind of research can only be conducted in a proper way if the textual data is in place, not only for the research proper, but also for those who want to reproduce this research later on. Therefore we base our research on the Amsterdam Hebrew Text Database (Van Peursen et al., 2015). This database is Open Access and can be downloaded from DANS, a national research archive in the Netherlands. Without downloading, the material in the database can be accessed through the website SHEBANQ (https://shebanq.ancient-data.org and Roorda, 2015b). Here the text of the Codex Leningradensis can be browsed, and while doing so, the user has access to a wealth of annotations that represent linguistic information and additional observations. In particular, there is an extensive set of cross-reference annotations between virtually all parallel passages (Roorda, 2015a). With the help of clustering techniques and entropy calculations we are investigating the challenge of linguistic variation in Biblical Hebrew and the transmission of the biblical texts. Results of others (Hornkohl, 2014:218-38, Rezetko and Young, 2014:380-83) and ourselves show that this approach leads to significant progress in our understanding of these issues.

The relevance of this research for digital humanities in general is that it explores challenges such as working with ancient languages of which we have only fragmentary evidence and (religious) texts with a long history of composition and transmission. While there is a lot of literature on these issues in traditional studies, it is clear that digital methods of research have not been pursued to their full potential yet.

Bibliography

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution, *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.*

Garcia, A. M. and Martin, J. C. (2007). Function Words in Authorship Attribution Studies, *Literary and Linguistic Computing*, **22**(1): 49-66.

Hornkohl, A. D. (2014). *Ancient Hebrew Periodization and the Language of the Book of Jeremiah: the Case for a Sixth-Century Date of Composition*, Leiden: Brill.

Hurvitz, A. (2013). *Late Biblical Hebrew*, Khan.

Khan, G. (ed.) (2013). *Encyclopedia of Hebrew Language and Linguistics*, Vol. 4, Leiden, Brill, 2013.

Kutscher, E. Y. (1974). *The Language and Linguistic Background of the Isaiah Scroll (1QIsaa),* STDJ 6. Leiden, Brill.

Oosting, R., Dyk, J. and Glanz, O., Valence Patterns of Motion Verbs, Semantics, Syntax and Linguistic Variation, to be published.

Roorda, D. (2015a). Parallel Passages, https://shebanq.ancient-data.org/tools?goto=parallel

Roorda, D. (2015b). The Hebrew Bible as Data: Laboratory - Sharing – Experience, http://arxiv.org/abs/1501.01866

Saenz Badillos, A. (2004). *A History of the Hebrew Language*, Cambridge: Cambridge University Press.

Segarra, S., Eisen, E. and Ribeiro, A. (2013). Authorship Attribution Using Function Words Adjacency Networks, *Proc. Int. Conf. Acoustics Speech Signal Processing*: 5563-5567.

SHEBANQ, https://shebanq.ancient-data.org

Van Peursen, W. T., et al. (2015). *Hebrew Text Database ETCB-C4b*. DANS. http://dx.doi.org/10.17026/dans-z6y-skyh

Young, I., Rezetko, R. and Ehrensvärd, M. (2008). *Linguistic Dating of Biblical Texts*, 2 volumes, London: Equinox Publishing.

Entities as topic labels: improving topic interpretability and evaluability combining Entity Linking and Labeled LDA

Federico Nanni
federico.nanni8@unibo.it
Data and Web Science Group, University of Mannheim

Pablo Ruiz Fabo
pablo.ruiz.fabo@ens.fr
LATTICE Lab, École Normale Supérieure, France

Introduction

Humanities scholars have experimented with the potential of different text mining techniques for exploring large corpora, from co-occurrence-based methods to sequence-labeling algorithms (e.g. Named entity recognition). LDA topic modeling (Blei et al., 2003) has become one of the most employed approaches (Meeks and Weingart, 2012). Scholars have often remarked its potential for distant reading analyses (Milligan, 2012) and have assessed its reliability by, for example, using it for examining already well-known historical facts (Au Yeung, 2011). However, researchers have observed that topic modelling results are usually difficult to interpret (Schmidt, 2012). This limits the possibilities to evaluate topic modeling outputs (Chang et al., 2009).

In order to create a corpus exploration method providing topics that are easier to interpret than standard LDA topic models, we propose combining two techniques called Entity linking and Labeled LDA; we are not aware of literature combining these two techniques in the way we describe. Our method identifies in an ontology a series of descriptive labels for each document in a corpus. Then it generates a specific topic for each label. Having a direct relation between topics and labels makes interpretation easier; using an ontology as background knowledge limits label ambiguity. As our topics are described with a limited number of clear-cut labels, they promote interpretability, and this may help quantitative evaluation.

We illustrate the potential of the approach by applying it to define the most relevant topics addressed by each party in the European Parliament's fifth term (1999-2004).

The structure of our work is as follows: We first describe the basic technologies considered. We then describe our approach combining Entity Linking and Labeled LDA. Based on the European Parliament corpus (Koehn, 2005),[1] we show how the results of the combined approach are easier to interpret or evaluate than results for Standard LDA.

Basic technologies

Entity Linking

Entity linking (Rao et al., 2013) tags textual mentions with an entity from a knowledge base like DBpedia (Auer et al., 2007). Mentions can be ambiguous, and the challenge is to choose the entity that most closely reflects the sense of the mention in context. For instance, in the expression Clinton Sanders debate, Clinton is more likely to refer to DBpedia entity Hillary_Clinton than to Bill_Clinton. However, in the expression Clinton vs. Bush debate, the mention Clinton is more likely to refer to Bill_Clinton. An entity linking tool is able to disambiguate mentions taking into account their context, among other factors.

LDA Topic Modeling

Topic modeling is arguably one of most popular text mining techniques in digital humanities (Brauer and Fridlund, 2013). It addresses a common research need, as it can identify the most important topics in a collection of documents, and how these topics are distributed across the documents in the collection. The method's unsupervised nature makes it attractive for large corpora.

However, topic modeling does not always yield satisfactory results. The topics obtained are usually difficult to interpret (Schmidt, 2012, among others). Each topic is presented as a list of words. It generally depends on the intuitions of the researcher how to interpret these tokens in order to propose concepts or issues that these lists of words represent.

Labeled LDA

An extension of LDA topic model is Labeled LDA (Ramage et al., 2009). If each document in a corpus is described by a set of tags (e.g. a newspaper archive with articles tagged for areas like "economics", "foreign policy", etc.), Labeled LDA will identify the relation between LDA topics, documents and tags, and the output will consist of a list of labeled topics.

Our approach

Labeled LDA has shown its potential for fine grained topic modeling (e.g. Zirn and Stuckenschmidt, 2014). The method requires a corpus where documents are annotated with tags describing their content. Several methods can be applied to automatically generating tags, e.g. keyphrase-extraction (Kim et al., 2010). Our source for tags is Entity linking. Since entity linking provides a unique label for sets of topically-related expressions across a corpus' documents, it can help researchers get an overview of different concepts present in the corpus, even if the concepts are conveyed by different expressions in different documents.

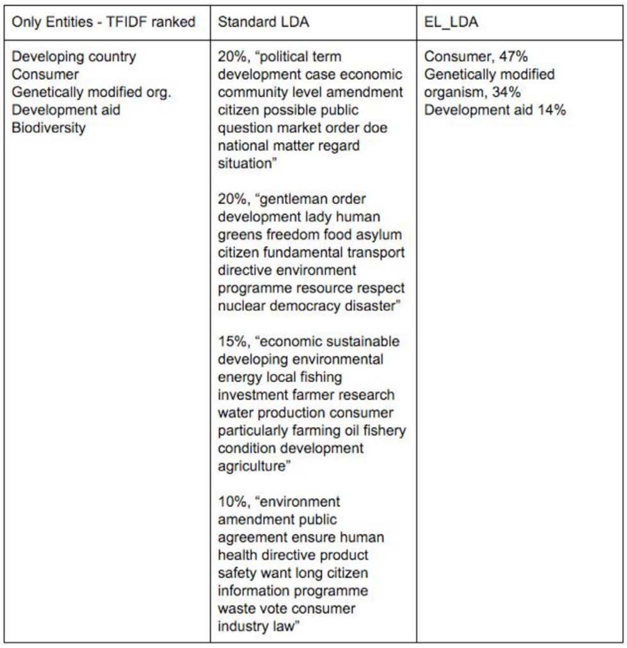
Our first step is identifying potential topic labels via entity linking. Linked entities were obtained with DBpedia Spotlight (Mendes et al., 2011). Spotlight disambiguates against DBpedia, outputting a confidence value for each annotation.[2] Annotations whose confidence was below 0.1 were filtered out. We also removed too general or too frequent entities (e.g. Country or European_Union)

We then rank entities' relevance per document with tf-idf (Jones, 1972), which promotes entities that are salient in a specific subset of corpus documents rather than frequent overall in the corpus. Finally, we select the top five entities per document as per tf-idf. These five entities are used as labels to identify, with Labeled LDA, the distribution of labeled topics in the corpus.

Experiments and Results

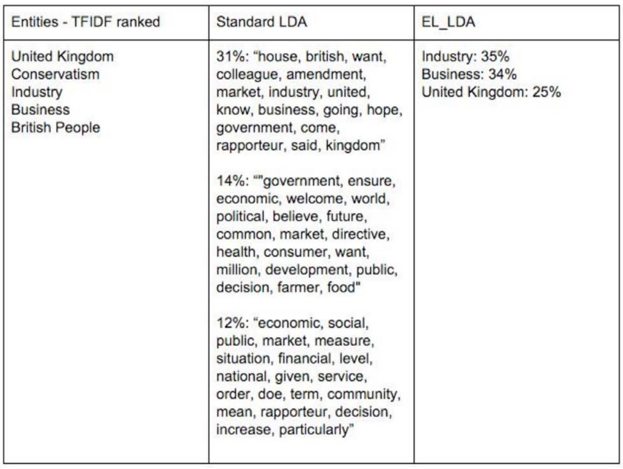
Using the Stanford Topic Modeling Toolbox,[3] we performed both Standard LDA (k=300) and Labeled LDA (with 5 labels)[4] on speech transcripts for the 125 parties at the European Parliament (1999-2004 session). The corpus contains 125 documents, representing one party each. Documents were tokenized and lemmatised; stop-words were removed. DBpedia entities were detected with Spotlight and ranked by tf-idf, as described above.

We present the outputs of Labeled LDA with entity labels (EL_LDA) for three parties, compared to both Standard LDA and to the top-ranked entities for each party (by tf-idf). In each case, we show topics with relevance above 10%. Results for the remaining parties are available online.[5]



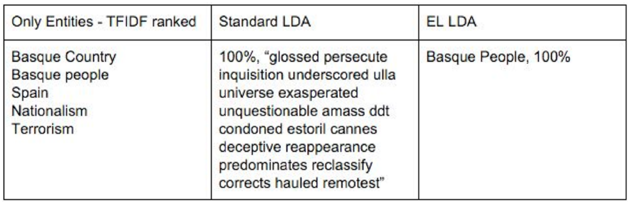
Only Entities - TFIDF ranked	Standard LDA	EL_LDA
Developing country Consumer Genetically modified org. Development aid Biodiversity	20%, "political term development case economic community level amendment citizen possible public question market order doe national matter regard situation" 20%, "gentleman order development lady human greens freedom food asylum citizen fundamental transport directive environment programme resource respect nuclear democracy disaster" 15%, "economic sustainable developing environmental energy local fishing investment farmer research water production consumer particularly farming oil fishery condition development agriculture" 10%, "environment amendment public agreement ensure human health directive product safety want long citizen information programme waste vote consumer industry law"	Consumer, 47% Genetically modified organism, 34% Development aid 14%

Figure: Linked entities (tf-idf-ranked), standard LDA topics and EL-LDA topics for speeches by Les Verts (France).



Entities - TFIDF ranked	Standard LDA	EL_LDA
United Kingdom Conservatism Industry Business British People	31%: "house, british, want, colleague, amendment, market, industry, united, know, business, going, hope, government, come, rapporteur, said, kingdom" 14%: ""government, ensure, economic, welcome, world, political, believe, future, common, market, directive, health, consumer, want, million, development, public, decision, farmer, food" 12%: "economic, social, public, market, measure, situation, financial, level, national, given, service, order, doe, term, community, mean, rapporteur, decision, increase, particularly"	Industry: 35% Business: 34% United Kingdom: 25%

Figure: Linked entities (tf-idf-ranked), standard LDA topics and EL-LDA topics for speeches by the Conservative Party (UK).



Only Entities - TFIDF ranked	Standard LDA	EL LDA
Basque Country Basque people Spain Nationalism Terrorism	100%, "glossed persecute inquisition underscored ulla universe exasperated unquestionable amass ddt condoned estoril cannes deceptive reappearance predominates reclassify corrects hauled remotest"	Basque People, 100%

Figure: Linked entities (tf-idf-ranked), standard LDA topics and EL-LDA topics for speeches by Partido Nacionalista Vasco (Spain).

Discussion

Labeled LDA combines the strengths of Entity Linking and standard LDA. Entity Linking provides clear labels, but no notion of the proportion of the document that is related to the entity. Standard LDA's relevance scores do provide an estimate to what an extent the topic is relevant for the document, but the topics are not expressed with clear labels. Labeled LDA provides both clear labels, and a quantification of the extent to which the label covers the document's content.

An advantage of Labeled LDA over Standard LDA is topic interpretability. Consider the UK Conservative Party's topics. In each standard LDA topic, there are words related to the concepts of *Industry* and *Business* in general, and some words related to the UK appear on the first topic. However, in each topic, some other words (e.g. *government, directive, decision, measure, health, consumer*) are related to other concepts, like perhaps *Legislation* or *Social policy*. A researcher trying to understand the standard LDA topics is faced with choosing which lexical areas are most representative of each topic: is it the ones related to *Industry*, *Business*, and the UK, or is it the other ones? The clear-cut labels from Labeled LDA are more interpretable than a collection of words representing a topic.

The Labeled LDA topics may be more or less correct, just like Standard LDA topics. But we find it easier to evaluate a topic via questions like "is this document about *Industry*, *Business* and *the UK*, in the proportions indicated by our outputs?" than via questions like "is this document about issues like *house, british, amendment, market, industry, government,* (and so on for the remaining topics)"?

The topics for French party Les Verts illustrate Labeled LDA's strengths further. Most of the Standard LDA topics contain some words indicative of the party's concerns (e.g. *environment* or *development*). However, it is not easy to point out which specific issues the party addresses. In Labeled LDA, concrete issues come out, like *Genetically modified organism*.

Topic label *Development aid* shows a challenge with entity linking as a source of labels. Occurrences of the word *development* have been disambiguated towards the entity *Development_aid*, whereas the correct entity is likely *Sustainable_development*. These errors do not undermine the method's usefulness. Efficient ways to filter out such errors exist; this is conceptually similar to removing irrelevant words from Standard LDA topics. However, we need to be aware of and address this challenge.

Regarding Partido Nacionalista Vasco (Basque Nationalist Party), the Standard LDA topic misses the word *basque*, which is essential to this party. Labeled LDA identifies *Basque people* as a dominant concept in this party's interventions.

Outlook

Our method performs Labeled LDA using Entity Linking outputs as labels. Its main advantage is providing a specific label for each topic, that improves topic interpretability, and can simplify human evaluation of topic models.

More evaluation is needed to fully assess the approach. We will consider two possible complementary evaluations: first, a crowdsourced task where participants evaluate the coherence of Labeled LDA topics with the corpus documents. Second, an assessment of our topics by political science experts. We're mostly interested in evaluating the approach for diachronic comparisons.

Bibliography

Au Yeung, C. M. and Jatowt, A. (2011). Studying how the past is remembered: towards computational history through large scale text mining. *Proceedings of the 20th ACM international conference on Information and knowledge management.*

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Berlin Heidelberg: Springer.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning research*, **3**: 993-1022.

Brauer, R., and Fridlund, M. (2013). Historicizing Topic Models, A distant reading of topic modeling texts within historical studies. *International Conference on Cultural Research in the context of "Digital Humanities"*, St. Petersburg: Russian State Herzen University.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems.*

Cornolti, M., Ferragina, P. and Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

Kim, S. N., Medelyan, O., Kan, M. Y. and Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit*.

Mendes, P. N., Jakob, M., García-Silva, and Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. *Proceedings of the 7th International Conference on Semantic Systems*. ACM.

Meeks, E. and Weingart, S. B. (2013). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, **2**(1): 2-1.

Milligan, I. (2012). Mining the "Internet Graveyard": Rethinking the Historians' Toolkit.> *Journal of the Canadian Historical Association/Revue de la Société historique du Canada*, *23*(2), 21-64.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rao, D., McNamee, P. and Dredze, M. (2013). Entity linking:

Finding extracted entities in a knowledge base. *Multi-source, Multilingual Information Extraction and Summarization*. Springer Berlin Heidelberg.

Salton, G., Fox, E. A. and Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM*, **26**(11): 1022-1036.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, **2**(1): 49-65.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **28**(1): 11-21.

Usbeck, R., Röder, M. and Ngonga, A. C. N. (2015). Evaluating Entity Annotators Using GERBIL. *The Semantic Web: ESWC 2015 Satellite Events*. Springer International Publishing.

Zirn, C. and Stuckenschmidt, H. (2014). Multidimensional topic analysis in political texts. *Data and Knowledge Engineering*, **90**: 38-53.

Notes

[1] http://www.statmt.org/europarl/

[2] Spotlight outperforms other systems when corpus entities often correspond to common-noun mentions like *democracy*, vs. proper-noun mentions (e.g. *Greenpeace*). See Cornolti et al., 2013 and Usbeck et al., 2015.

[3] http://nlp.stanford.edu/software/tmt/tmt-0.4/

[4] Each document (party) is labeled with 5 entities. Some entities are shared across parties. For the 125 parties, this gives 300 distinct labels. This corresponds to k=300 topics in Standard LDA.

[5] https://sites.google.com/site/entitylabeledlda

Visualising Cultural Spheres – Virtual Tours and Epigraphical Data

Anna Neovesky
anna.neovesky@adwmainz.de
Academy of Sciences and Literature | Mainz, Germany

Max Grüntgens
max.gruentgens@adwmainz.de
Academy of Sciences and Literature | Mainz, Germany

Introduction

Today, cultural heritage sites, museums and other places of historical or societal value can often be visited on the Internet. Panoramic images and virtual tours allow the user to access distant sites from home via handheld devices as well as conventional desktop devices. In this way, these applications strongly reduce the threshold for getting acquainted with various cultures, their respective artefacts and unique heritage.

But can this popular and usually touristic way of presentation be used to introduce valid scientific information to a broad public? This question has been posed at the Academy of Sciences and Literature | Mainz regarding its project "Die Deutschen Inschriften".

The research project "Die Deutschen Inschriften"

The long term research project "Die Deutschen Inschriften" is a joint undertaking of six German Academies of Sciences and the Austrian Academy of Sciences. The research focuses on collecting, editing and interpreting medieval and early modern Latin and German inscriptions. They often occur in conjunction with figurative elements or spatial as well as architectural features. The inscriptions themselves are mostly in medieval Latin or in historical or regional varieties of the German language. The geographical area of research consists of Germany, Austria and South Tyrol. The inscription records range from approximately 500 AD to 1650 AD (Brandi, 1937; Kloos, 1973; Nikitsch, 2008). The project's scholars carry out their research within a wide scope of interests ranging from art history, philology and linguistics to the history of ideas. The research results are published in 90 volumes. More than 43 of these volumes, including over 17.000 records, are currently accessible through the online database "Deutsche Inschriften Online" (German Inscriptions Online, www.inschriften.net).

Virtual cultural heritage

Observing an item within a cultural heritage site in isolation frequently limits the understanding of it. This is due to its removal from the big picture of the entire ensemble in its historical, cultural and spatial context.

Two different approaches of representing historical sources in their spatial context are being explored by the projects "Inschriften im Bezugssystem des Raumes" (Inscriptions in their Spatial Context, IBR) and the virtual tours through St. Stephan in Mainz and St. Michael in Hildesheim. Project IBR utilised methods of laser scanning and semantic web technologies, in this regard aiming at a more specialised target audience. The virtual tours of St. Stephan and St. Michael on the other hand were developed as a means to visualise the spatial cultural sphere for an audience with a lower degree of specialized knowledge. In doing so the applications were generally aiming at a broader audience (Lange/Unold 2015; www.spatialhumanities.de/ibr/startseite.html; www.inschriften.net/hildesheim/ rundgang.html)

The virtual tour's objective was to arrange the scientific edition's epigraphical items in their spatial context and to put the scientific sources on display to a diverse audience in an easy accessible and comprehensible manner.

In Hildesheim, the interrelation between the church and bishop at the time of construction, Bernward of Hildesheim, is shown and emphasised through the inscriptions and their placement. (Kruse, 2012). So, the visualisation of this connection and the importance of inscriptions as biographical evidence as well as general historic sources are a further aim.

Generic virtual tours

Several software applications for creating virtual tours already exist that enable everyone to build virtual tours. The software offers the modelling of the views, their arrangement within a tour, the navigation, interactive elements as well as the display on a map. Images, short text and links can be added to describe selected objects. The lack of possibilities to import and embed detailed textual information and to load them from external repositories, limits the benefits in a scientific context. Thus, a software for generic virtual tours was developed to integrate and provide a generic approach for adding content from external repositories (databases, text documents, etc.).

Taking the tour to St. Michael in Hildesheim as an example for the possibilities: the virtual tour enables the user to navigate within the church accessing several fixed viewpoints, to look around and to zoom in on interesting looking spots. Interactive elements indicate the possibility of switching to a different viewpoint, provide information about specific points of interest, most of them inscriptions on stone tablets or tomb slabs or other (art-)historic artefacts. Pop-up windows can contain general information about the inscriptions, a full transcription, as well as a translation of the inscribed texts. Furthermore, images of the inscription, historical sketches, or old photographs as well as audio information can be provided. All this information is displayed in the context of the tour, without exiting the panoramic visualisation. Links to the critical edition of the inscriptions in “Deutsche Inschriften Online” provide easy access to the full scholarly content including scientific apparatus and further bibliographic introductory material.

The software allows for subdividing the information about the epigraphical and pictorial agenda utilized throughout the church into multiple sub-tours. Each tour is therefore enabled to concentrate on a specific topic and by this means to weave a unique narrative of the site and its cultural sphere. Furthermore, various preservation stages and relocations of objects and decorations can be pointed out. The conception of such thematic tours prevents the user from "getting lost" in the virtual environment and from overlooking the content's contextual conclusions and messages. Throughout the tour the viewer receives information in structured order, e.g. the events are sorted chronologically as is the rule in the context of historical information (Rizvic, 2014; Tan/Rahman, 2009).

Cultural coding = generic coding

Creating digital code in the public sector is by necessity an open source process. In contrast to coding in the realm of the competitive private sector, in the cultural realm it is not—and must not be—imperative to shield digital products, project data and know how. Open source software enables people from outside the project to reuse and adapt the outcome as well as to contribute. A generic approach increases the reusability of the code and the application.

Hence the virtual tour was designed as browser-based and implemented as a generic JavaScript application using HTML5, WebGL and the Three.js framework. (Neovesky/Peinelt, 2015). An encapsulated package was released that enables others to build their own virtual tour without programming knowledge. A simple JSON file is used for configuration and data exchange. Adaption of this core JSON file makes it possible to create a custom virtual tour. All the cultural institution has to add are suitable images, positions and texts which can for example be received via data interfaces and web APIs. The generic virtual tour and the user manual are available on GitHub (https://github.com/digicademy/virtualTour).

Conclusion

The virtual tour to the inscriptions of St. Michael’s Church in Hildesheim was born from the idea of combining a popular visual representation with unrestricted access to scientific research and publications.

By means of using a generic approach to the technical implementation, the software is not limited to any specific set of circumstances and does not represent a single-use scenario. Although the application was developed with a specific object in mind, it can easily be adapted to fit other projects within the fields of research, education and visualisation.

Even though the application focuses upon a broad audience, the visualisation can also be of interest to researchers from diverse academic background and disciplines. In this manner scholars are able to receive an impression of a distant site’s or compound’s spatial context. They are as well enabled to present their research in an easy to use, appealing and low-threshold way. The distribution under a free license (GNU GPL V3) enables users to use, adapt and extend the software.

Bibliography

Brandi, K. (1937). Grundlegung einer deutschen Inschriftenkunde. *Deutsches Archiv für Erforschung des Mittelalters Bd. 1*. Münster/Köln: Böhlau-Verlag, pp. 11–43, URL: http://www.digizeitschriften.de/dms/img/?PID=PPN345858700_0001%7Clog10 (accessed 4th March 2015).

Kloos, R. M. (1973). Die Deutschen Inschriften. Ein Bericht über das deutsche Inschriftenunternehmen. *Studi medievali, 3a serie*, Spoleto: Centro, Vol. **14**, pp. 335–62.

Kruse, K. (2012). Zur Bautätigkeit Bischof Bernhards in Hildesheim. In Lutz, G. and Weyer, A. (Eds.), *1000 Jahre St. Michael in Hildesheim*. Petersberg: Imhof, pp. 41–65.

Lange, F. and Unold, M. (2015). Semantisch angereicherte 3D-Messdaten von Kirchenräumen als Quellen für die geschichtswissenschaftliche Forschung. In Baum, C. and Stäcker, T. (Eds.), *Grenzen und Möglichkeiten der Digital Humanities (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1)*. DOI: 10.17175/ sb001_015 (accessed 4th March 2015).

Neovesky, A. and Peinelt, J. (2015). A Virtual Tour to the Inscriptions of the UNESCO World Heritage Site St. Michael in Hildesheim. *Electronic Visualisation and the Arts (EVA 2015) Conference Proceedings, London*. DOI: 10.14236/ewic/eva2015.31 (accessed 4th March 2015).

Nikitsch, E. J. (2008). Fritz V. Arens als Mainzer Inschriftensammler und Epigraphiker. *Mainzer Zeitschrift* **103**, pp. 231–43.

Rizvic, S. (2014). Story Guided Virtual Cultural Heritage Applications. *Journal of Interactive Humanities,* **2***(1),* DOI: 10.14448/jih.02.0002 (accessed 4th March 2015).

Tan, B.-K. and Rahman, H. (2009). *Virtual heritage: reality and criticism. In CAADFutures 2009: Joining Languages, Cultures and Visions, Montreal*, pp. 143–56.

Stefan George Digital: Exploring Typography In A Digital Scholarly Edition

Frederike Neuber
frederike.neuber@uni-graz.at
Graz University, Austria

Introduction

The digital scholarly edition *Stefan George Digital* (StGD) is, as its name implies, dedicated to the oeuvre of the German author Stefan George (1868-1933).[1] As part of my PhD thesis, StGD is concerned with the role and application of typography within the printed collections of George's poetry and the formal canon, and development of the so-called *Stefan-George-typeface (St-G-typeface)*. To capture typographical information within the digital edition, I have tested two approaches that this short paper will discuss: the application of a TEI-based model and the development and the integration of an ontology.

Subject matter

More than any other poet in modern German literature, Stefan George (1868-1933) engaged with writing – particularly typography[2] – in exceptional ways. From 1897 on, he almost completely abandoned his cursive handwriting, using instead highly stylized block letters. From 1904 on, this individual book hand was transferred into metal:[3] the third edition of *Das Jahr der Seele* (*The Year of the Soul*) was the first volume printed in the so called *St-G-typeface*, a Sans-Serif typeface which emerged when the German dispute between Serif and Black Letter typefaces was in full swing. Accordingly, St-G forms a third, alternative typeface, strongly inspired by modern Sans-Serif typefaces such as the Akzidenz-Grotesk of the Berthold foundry. Furthermore, St-G includes letter shapes of Roman and Carolingian scripts as well as of the Greek alphabet (Kurz, 2007). Between its inception and 1927, the *St-G typeface* was changed several times, so that it now exists in various versions.

Problem statement and project goals

George linked his poetry and his understanding of aesthetics to the design of his books by introducing an individualized typeface. He broke with typographical conventions at the time by applying a Sans-Serif typeface, by basing the design on his own handwriting, and by referencing historical script models in its formal features. The extraordinary design of St-G and the fact that the author himself was involved in its creation calls for special attention in a scholarly edition. However, previous editions neither include a detailed recording of the printed publications nor their typographical analysis. StGD aims to close this gap by providing a digital scholarly edition that allows for exploring typography in George's poetical work.

In the first phase of research, I will create a digital edition of printed poetry collections by George. I will develop a model to identify and describe typographical forms as well as to allow for citing them. In a second phase of research I will enhance the corpus of StGD with handwritten drafts, thereby allowing further investigations of the relationship between George's book hand and the typeface.[4]

The digital edition

The corpus of StGD consists of 29 printed editions of Stefan George's poetical works published between 1890 and 1933. Individual works are represented in the corpus variable numbers of times according to their textual and typographical variation.[5] Currently all volumes are encoded according to a customized XML/TEI schema and enriched by bibliographical metadata through FRBR (Functional Requirements for Bibliographic Records) and METS (Metadata Encoding and Transmission Standards). The full texts will be enhanced by corresponding facsimiles, provided via a IIIF (International Image Interoperability Framework) compliant image server using the OpenSeadragon viewer. At the end of the project (April 2017), all contents of the digital edition will be openly available through a Creative Commons (CC-BY-NC-SA) License via the FEDORA-based asset management system

GAMS (Geisteswissenschaftliches Asset Management System)[6].

Focus of the paper

Due to the significant role of typography in George's work, the creation of a digital scholarly edition calls for special attention to graphical features within the documents. This means that typography needs to be classified (i.e. typeface family, font) and its features need to be modelled. The latter includes the description of typographical forms, the identification of stylistic models, and the definition of the semantic function of types in the text. Such typographical enrichment is particularly challenging since neither a commonly shared vocabulary to describe typographical features nor a widely accepted type classification system exists. With regard to this lack of a common standard, StGD has mainly tested out two methods of typographical enrichment which will be discussed in this paper: the application of a TEI based model and the development of an ontology to describe typography.

Typography and the TEI

A distinction can be drawn between two different purposes of encoding features of writing: representation and information enrichment. The first purpose is covered to a great extent by the TEI gaiji module[7] and the application of Unicode. Recent editing projects like *Hugo von Montfort: the poetical work*[8] demonstrate the potentials of these methods, even if they also show that performing them throughout a complete edition is work-intensive and impractical. Concerning the capture of information about writing, especially about typography, the possibilities offered by the TEI are more restricted. Although the element <typeDesc>[9] as part of <msDesc> allows for a description of types in prose, there is no TEI vocabulary to describe types and their features in a formalized way. The paper will report on the benefits and drawbacks of the already implemented elements and attributes and demonstrate their application to material at hand.

Typography and Ontologies

There are barely any digital projects dedicated to the modelling of typography. Those that exist include the *Type Repository of Incunabula*[10] at the Berlin State Library, a database which identifies and catalogues incunabula types. It applies a relatively flat project-customized XML schema that describes types in prose. However, the digital modelling of handwriting has made significant progress over the last decade, since Arianna Ciula coined the term "Digital Palaeography" (Ciula, 2005), and the modelling of typography can benefit from this research. Recent research projects like *DigiPal*[11] and *ORIFLAMMS*[12], though not created in the context of digital scholarly editions, emphasize the strong tendency towards the application of semantic web technologies for the modelling of handwriting (Stokes, 2011 and 2012; Stutzmann, 2013). This paper will discuss their advantages for the identification, formal description and citation of typographical forms. Moreover it will give an overview of the technology or modelling language (i.e. RDFs, UML, SKOS) that might be the most suitable for the purpose of StGD. In this context, the paper will strongly take into account aspects of practicability and re-usability of the model.

Research questions

By modelling and analysing typographical information, the digital edition opens up Stefan George's poetical work for the following research questions: (a) Which formal features does the St-G-typeface contain and what are they referring to?; (b) How has the formal canon of the typeface developed between 1904 and 1927?; (c) Is the development of the book design across George's work linear or is it marked by any significant breaks?; and (d) How are text and typography interrelated in George's work, and how is typography applied as a stylistic device?

Context in Digital Humanities

StGD is contributing to the field of digital scholarly editing, the focus of which is shifting increasingly to the materiality of the edited documents. This tendency has been encouraged by movements such as New Philology as well as by textual concepts such as the "material text: (Shillingsburg 1997) and the idea of text as interaction of "bibliographical" and "linguistic code" (McGann 1991). In this context writing – as the interface between the text's message and its documentary carrier – plays a significant role and is particularly difficult to capture (Schubert, 2010). The goal of StGD is to aid in the development and promotion of a future best practice method for the modelling of typography in digital scholarly editions. Furthermore, the project's thematic focus contributes to the field of digital book history. By putting typography in its focus, StGD represents a first step towards the burgeoning and as-yet-unexplored field of Digital Studies of Typography.

Bibliography

Ciula, A. (2005). *Digital palaeography: using the digital representation of medieval script to support palaeographic analysis.* Digital Medievalist 1, ed. by D. P. O'Donnell et al. <http://www.digitalmedievalist.org/journal/1.1/ciula/>. (all URLs accessed 6 March 2016).

Kurz, S. (2007). *Der Teppich der Schrift.* Frankfurt/M.: Stroemfeld.

McGann, J. J. (1991). *The textual condition.* Princeton University Press.

Pierazzo, E. (2015). *Digital Scholarly Editions. Theories, Models and Methods.* Aldershot: Ashgate.

Shillingsburg, P. (1997). *Resisting Texts. Authority and Submis-*

sion in Constructions of Meaning. Ann Arbor: University of Michigan Press.

Schubert, M. (2010). *Einleitung, in Materialität in der Editionswissenschaft*, ed. by M. Schubert. Berlin, New York: De Gruyter, pp. 1-13.

Stokes, P. (2011). *Describing Handwriting*, Part I [and following]. London. <http://www.digipal.eu/blogs/tag/describing-handwriting/>

Stokes, P. (2012). *Modeling Medieval Handwriting: A New Approach to Digital Palaeography.* DH2012 Book of Abstracts, ed. by Jan Christoph Meister et al. Hamburg, pp. 382–85. <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-medieval-handwriting-a-new-approach-to-digital-palaeography>

Stutzmann, D. (2013). *Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS.* Document numérique, **16**, (Ed). by Christine Bénévent, et al. Paris: Lavoisier (Hermes), pp. 81-96.

Wehde, S. (2000). *Typographische Kultur: Eine zeichentheoretische und kulturgeschichtliche Studie zur Typographie und ihrer Entwicklung.* Berlin, New York: De Gruyter.

Notes

[1] The digital edition Stefan George Digital is developed in the context of a DiXiT (Digital Scholarly Editions Initial Training Network) fellowship, funded under Marie Curie Actions within the European Commission's 7th Framework Program.

[2] Although the term typography includes both the micro- and macro-design of a print (and both are equally important when engaging with George's work), this paper will focus solely on the level of micro-typography, by which I mean the level of the choice, design, and arrangement of types.

[3] It is still unclear who designed the St-G typeface, though it is presumed that both the book designer Melchior Lechter and the printer Otto von Holten were involved.

[4] The second research phase is planned in the context of a five-months visiting fellowship within the *DigiPal* project at King's College London in Spring 2016. Presumably the initial results of this collaboration will also be presented shortly in this paper.

[5] i.e. *The Tapestry of Life (Der Teppich des Lebens)* is represented four times (1900, 1901, 1904, 1932), *The Star of the Convenant (Der Stern des Bundes)* two times (1914, 1928).

[6] <http://gams.uni-graz.at/> [all quoted URLs accessed 6.3.2016]

[7] <http://www.tei-c.org/release/doc/tei-p5-doc/de/html/WD.html#D25-20>

[8] Website of the digital edition *Hugo von Montfort: the poetical work*: < http://gams.uni-graz.at/collection:me>; example of a XML/TEI encoding: <http://gams.uni-graz.at/archive/objects/o:me.1r/datastreams/TEI_SOURCE/content>.

[9] <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-typeDesc.html>

[10] Website of the *Type Repository of Incunabula*: <http://tw.staatsbibliothek-berlin.de/>; example of XML encoding: <http://tw.staatsbibliothek-berlin.de/materials/ma04679.xml?_xsl=no>

[11] http://www.digipal.eu/>

[12] <http://www.agence-nationale-recherche.fr/?Project=ANR-12-CORP-0010>

A Comparative Analysis of Bibliographic Ontologies: Implications for Digital Humanities

Terhi Nurmikko-Fuller
terhi.nurmikko-fuller@oerc.ox.ac.uk
University of Oxford, United Kingdom

Jacob Jett
jjett2@illinois.edu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Timothy Cole
t-cole3@illinois.edu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Chris Maden
crism@illinois.edu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Kevin R. Page
kevin.page@oerc.ox.ac.uk
University of Oxford, United Kingdom

J. Stephen Downie
jdownie@illinois.edu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Introduction

In the ever-expanding world of digital libraries and cultural heritage collections, bibliographic metadata standards provide a structured approach to managing resources. The capture of additional contextual information further supports the identification, selection, and use of the resources described. As the problem spaces and areas of study of Humanities scholars are increasingly diversified (Henry and Smith, 2010) and supported by digital material and methods, existent approaches and systems are falling short of the needs of users (Fenlon et al., 2014; Varvel and Thomer, 2011). In short, research agendas and investigations have begun to evolve beyond searches based on traditional metadata parameters (author, date, publication place, genre).

Semantic Web technologies have been identified as a potential solution (Bair and Carlson, 2008). They augment keyword-based, full-text approaches to discovery, with methods that rely on named entity identification, relationships between entities, and the potential to leverage interlinked data from a variety of repositories and corpora. A number of different, well-defined ontologies (structural frameworks capturing domain information) created by various bodies within the wider context of the Digital Humanities (DH) have emerged (Isaac, 2013; Stead, 2006). A critical evaluation and comparison between these different structures has, however, been lacking.

In this paper, we provide a summary of four bibliographical metadata ontologies, and expand on an earlier initial comparative analysis between them. What follows is a more in-depth discussion of complementary differences and parallels in terms of expressiveness, rather than domain, focus, or perspective. The strengths and weaknesses of these vocabularies are of interest and importance to anyone working with any type of Humanities dataset or research output, whether it be interacting with metadata that describes the resource, features that have been extracted from it, or the resource itself.

Bibliographic Ontologies

To date digital libraries have relied heavily on traditional library bibliographic standards, such as MARC[1]. As new research questions have arisen in DH, the limitations of earlier standards have become more pronounced (see Ramesh et al., 2015; Sfakakis and Kapidakis, 2009; Park, 2006; Cantara, 2006; Shreeves et al., 2005), and a number of ontologies designed to map the entities and relationships inherent in bibliographical metadata have emerged.

Rather than aiming to provide a comprehensive analysis of all known examples, we extended a preliminary evaluation of a small number of bibliographic ontologies. Earlier research[2] bridging the large general corpus of the HathiTrust Digital Library[3] with the specialist Early English Books Online - Text Creation Partnership[4] assessed the different needs of three distinct case study examples and analysed the suitability of existing ontologies to adequate capture associated information, including publication facts and object biographies (Nurmikko-Fuller et al., 2015a). This preliminary analysis examined four ontologies: MODS RDF[5]/ MADS RDF[6], BIBFRAME (Miller, et al., 2012), Schema.org[7], and FRBRoo (Bekiari, et al., 2013). BiBo[8] was originally considered, but excluded from the final comparison as it operates on a finer level of granularity.

In this paper, we elaborate on that initial analysis, and provide access to the entirety of the comparative table (Nurmikko-Fuller et al., 2015b)[9] of which only a representative sample has previously been made available. We summarise the main characteristics of each structure in order to provide context for the detailed discussion outlining the parallels and differences between the models.

Methodology

Based on available documentation and extant examples, we conducted an extensive review of the expressiveness of each ontology. Comparing each property and class against possible alignments in the other three led to the identification of parallels and differences between these models. One revelation was the differing extent to which documentation had been left incomplete, highlighting the lack of workflow standardisation in ontology-development even within a shared domain. At times the absence of extensive documentation complicated our ability to confidently assert parallels between the models.

The comparative analysis led to the insertion of all classes and properties of each ontology into one cohesive table, aligned wherever the same data could be represented regardless of how the mapping was achieved, and resulting in a table of exactly 500 rows. Five types of alignment were identified:

- equivalent, where the same data can be mapped in each ontology using a single class. An example of this is location information (such as publication place), captured via madsrdf:Geographic, bf:Place, sc:Place, and frbroo:F9_Place (equivalent to cidoc:E53) in MODS/MADS, BIBFRAME, Schema.org and FRBRoo respectively.
- alternative, which encompasses situations where properties were used in one ontology, but classes in another to talk about the same attribute. An example of this is Schema.org's sc:birthDate. It takes an entire grouping of entities and relationships to express this same information in FRBRoo (frbroo:P98B_wasBorn frbroo:E67_Birth frbroo:P4_hasTime-Span frbroo:E52_Time-Span frbroo:P78_isIdentifiedBy frbroo:E50_Date).
- parallel, where the same data can be mapped using a combination of classes and properties. In the case of the date of creation for a work, MODS/MADS, BIBFRAME and Schema.org all have a single property (mods:dateCreated, bf:creationDate, CreativeWork:dateCreated), whilst FRBRoo necessitates a cluster of classes and properties: F1_Work R19b_wasRealisedThrough F28_ExpressionCreation P2_hasType E55_Type{"Earliest known externalisation"}. We consider these approaches as aligned because the same data can be mapped against them, but parallel rather than exactly equivalent due to different approaches.
- partial, where the same data could be mapped using different ontologies to a greater or lesser extent. As an example of this, we cite the assignment of a unique identifiers, captured using a single property in MODS/MADS (modsrdf:identifier) and FRBRoo (which uses a CIDOC property cidoc:P48_hasPreferredIdentifier), but through several possible options in Schema.org (Thing:sameAs, Book:isbn, Periodical:issn), and via a total of 13 possible properties in BIBFRAME (such as bf:doi, bf:isbn, bf:uri).
- granular, which captures differences in levels of granularity. This is illustrated by entity types such sc:CreativeWork and sc:Book, and showcases how categorical alignment across all four ontologies is not always possible. In the case of frbr:F1_Work (a conceptual version of a work, of which digital and physical manifestation are carriers), an equivalent alignment can be drawn to bf:Work,

but MODSRDF/MADSRDF has no entity type that fulfills the role of representing that notion.

Comparative Analysis

The bibliographic metadata ontologies discussed here differ in their approach and expressiveness. Of the four, MODS RDF/MADS RDF was found to be most descriptive, with FRBRoo an event-based model, and BIBFRAME bridging the two by virtue of containing characteristics typical of either. Schema.org stands out as an ontology that promulgates a model at the crossroads between the other four; however, its focus on instrumenting marketplace transactions also detracts from much of its descriptive power and leaves it orthogonal to the purposes of the others. It has some generic properties that are useful in each of the other ontologies, but also possesses properties and classes that end users do not require outside of point of service systems.

From the perspective of a DH user of these ontologies, each is (to some degree) a victim of its provenance and the motivations of its designers. The benefits and failings of each are different, and they all incorporate a number of idiosyncrasies:

- MODS/MADS has a data structure that replicates the XML serialization format and is frequently realized in RDF as empty nodes.
- Schema.org affords humanists with a vocabulary that combines events and descriptions but differs from the other models in its focus.
- FRBRoo spans beyond the remit of bibliographical metadata by mapping relationships via temporal entities; this results in greater complexity for the representation of the same data, often necessitating a cluster of classes and properties.
- BIBFRAME adopts a different method, recreating MARC in RDF using methods and approaches more in line with graphical thinking, as well as extending beyond that format.

Conclusion

Our review examines the structure and scope of four ontologies designed for the representation of bibliographic metadata as applied to cataloguing digital source material in the Humanities. From this analysis direct equivalences, parallels, and complementary differences have emerged: there are many similarities in aim, scope, and expressiveness, but none of the considered ontologies completely satisfy scholarly needs on their own. Moving between them is feasible, but not achievable without some lossiness, as illustrated by the examples for granular alignment (see **Methodology**). For the comprehensive mapping of all the aspects of a given dataset, these models need to be supplemented with less bibliographically-focused ontologies. Our analysis has highlighted the need to formalize the mappings, best practices, and transformations, as these are key to the correct (re)use of ontologies across projects and domains.

We have provided DH researchers with a window into the digital corpora design process. Knowing the requirements of domain scholars to have interactions with finer-grained research objects, we will be looking at standards like BiBo, Web Annotation, and others during the next round of research.

Acknowledgement

The authors gratefully acknowledge their colleagues Pip Willcox, Bodleian Libraries, University of Oxford; Colleen Fallaw, and Megan Senseney, Graduate School of Library and Information Science University of Illinois at Urbana-Champaign, for their invaluable contributions to the creation of the Bibliographic Ontologies Comparative Features Dataset, available at https://www.ideals.illinois.edu/handle/2142/88356.

Bibliography

Bair, S. and Carlson, S. (2008). Where keywords fail: Using metadata to facilitate digital humanities scholarship. *Library Metadata* **8**(3): 249-62.

Bekiari, C., Doerr, M., Le Bœuf, P. and Riva, P. (2013). *FRBR object-oriented definition and mapping from FRBRER, FRAD and FRSAD* (version 2). International Working Group on FRBR and CIDOC CRM Harmonisation.

Cantera, L. (2006). Long-term preservation of digital humanities scholarship. *OCLC Systems & Services* **22**(1): 38-42.

Fenlon, K., Senseney, M., Green, H., Bhattacharyya, S., Willis, C. and Downie, J. S. (2014). Scholar-built collections: A study of user requirements for research in large-scale digital libraries. *Proceedings of the 77th ASIS&T Annual Meeting*, Seattle, WA, 31 October – 5 November 2014.

Henry, C. and Smith, K. (2010). Ghostlier demarcations: large-scale text digitization projects and their utility for contemporary humanities scholarship. *The idea of order : transforming research collections for 21st century scholarship*:106–115. Council on Library and Information Resources. http://www.clir.org/pubs/reports/reports/pub147/pub147.pdf (accessed March 3 2016).

Isaac, A. (ed.) (2013). *Europeana Data Model Primer*. http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf (accessed March 3 2016).

Miller, E., Ogbuji, U., Mueller, V. and MacDougall, K. (2012). *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services*. Report. Library of Congress.

Nurmikko-Fuller, T., Fallaw, C., Jett, J., Page, K. R., Cole, T. W., Maden, C., Senseney, M. and Downie, J. S. (2015a). Bibliographic Ontologies Comparative Features Dataset. Champaign, IL: University of Illinois. http://hdl.handle.net/2142/88356

Nurmikko-Fuller, T., Page, K. R., Willcox, P., Jett, J., Maden, C., Cole, T., Fallaw, C., Senseney, M. and Downie, J. S. (2015b). Building Complex Research Collections in Digital

Libraries: A Survey of Ontology Implications. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 21–25 June 2015. DOI= http://dx.doi.org/10.1145/2756406.2756944

Park, J.-R. (2006). Semantic interoperability and metadata quality: An analysis of metadata item records of digital image collections. *Knowledge Organization* **33**(1): 20-34.

Poulter, A. (2010). One ring to rule them all: CIDOC CRM. *Catalogue & Index* 161, pp. 31-33.

Ramesh, P., Vivekacardhan, J. and Bharathi, K. (2015). Metadata diversity, interoperability and resource discovery issues and challenges. *DESIDOC Journal of Library & Information Technology*, **35**(3): 193-99.

Sfakakis, M. and Kapidakis, S. (2009). Eliminating query failures in a work-centric library meta-search environment. *Library Hi Tech* **27**(2): 286-307.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B. and Cole, T. W. (2005). Is "quality" metadata "shareable" metadata? The implications of local metadata practices for federated collections. In H.A. Thompson (Ed.) *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries*, April 7-10 2005, Minneapolis, MN. Chicago, IL: Association of College and Research Libraries. pp. 223-37.

Stead, S. (2006). *The CIDOC CRM: a standard for the integration of cultural information*. http://www.cidoc-crm.org/cidoc_tutorial/index.html (accessed November 15 2010).

Varvel, V. E. J. and Thomer, A. (2011). *Google digital humanities awards recipient interviews report (CIRSS Report No. HTRC1101)*. Technical report prepared for the HathiTrust Digital Library. Champaign, IL: Center for Informatics Research in Science and Scholarship.

Notes

[1] http://www.loc.gov/marc/bibliographic/

[2] http://www.oerc.ox.ac.uk/projects/elephant

[3] https://www.hathitrust.org/htrc

[4] http://www.textcreationpartnership.org/tcp-eebo/

[5] http://www.loc.gov/standards/mods/modsrdf/v1/

[6] http://www.loc.gov/standards/mads/rdf/v1.html

[7] http://schema.org/docs/full.html

[8] http://bibliontology.com/

[9] http://hdl.handle.net/2142/88356

Digging the Aboveground: Visual Archeology of an Asylum

Burcak Ozludil Altin
bozludil@njit.edu
New Jersey Institute of Technology, United States of America

Augustus Wendell
wendell@njit.edu
New Jersey Institute of Technology, United States of America

Psychiatric practice during the nineteenth century was closely engrained in the spaces of the asylum. Envisioned as "therapeutic instruments" in and of themselves, asylums boomed all around the world: some were purpose-built, some were existing buildings repurposed as asylums. But how did these institutions really function? How was psychiatry practiced here?

Our proposal reports a work in progress that visualizes the evolution and workings of a nineteenth century Ottoman asylum. Located in Istanbul, the capital of the Ottoman Empire, it was used as the state mental hospital between 1873 and 1922. The building is originally a sixteenth century imperial complex by the famous Ottoman court architect Mimar Sinan (1489-1588) and is still standing albeit currently under heavy and damaging restoration. Research shows that during the time of its use as an asylum, the building changed significantly to enable its medical function. Indeed, it is precisely through the changing of the building one can demonstrate the medicalization and modernization of Ottoman psychiatry during this period. In this sense, our case study does not visualize an "ideal" asylum nor is it a search for the "original" complex; on the contrary, it analyzes a neglected and messy phase of an existing building's life. Traditionally perceived as a "corruption" in the original structure, we take the period it was used to treat the insane as a story worth being told.

The information is gathered from a detailed research of primary archival and printed sources involving multiple disciplines: medicine, psychiatry and architecture. These sources include: (1) Ottoman and Turkish official documents kept in the Prime Minister's Archives: Correspondences between state departments giving details about the running of mental institutions, spatial interventions, and concerns over public health; (2) Medical publications of the period: Visitor accounts by physicians, medical writings discussing approaches in psychiatry, and comparative reports on the current conditions of asylums around the world; (3) Popular media: Newspaper articles and travel accounts.

The project is located at the intersection of architectural and medical history; however, it is primarily using the toolset of architectural history. Partly stemming from the demands of the field and partly due to the nature of the project itself, one of the major concerns is the visualization of findings. In line with architects' particular understanding of visualization, the first resort was experimenting with various 2D and perspectival illustration techniques. 2D depictions (i.e. plan, façade, elevation) have been the mode of representation for centuries in the field of architecture. Perspective was a welcome addition with the development of linear, or mathematical, perspective in the fifteenth century. Computational 3D modeling and visual simulation (rendering) has become the norm starting from the 1990s. Despite its popularity in architectural practice and education, the use of 3D computation has, until recently, been relatively underutilized in architectural history.

Scholars in architectural history and in the neighbouring fields of art history and archeology have recently been using methodologies and approaches to integrate innovative technologies in historical research. From virtual reconstruction to web-based panoramas, geospatial modeling to photogrammetry, the virtual creation of static environments to interactive ones, these studies appear in multiple platforms and venues related to digital humanities, digital art history, architectural computation etc. In the mainstream publications of architectural history, an important milestone was the publication of an article that featured real-time interactive simulations of the Roman Forum presenting a hypothetical reconstruction of a funeral ceremony in the Journal of the Society of Architectural Historians (JSAH) (Favro and Johanson, 2010). Another outstanding project that addressed issues of time and movement was "The Virtual Monastery" that focused on a single structure to analyse the workings of a building type through ages (Bonde et al., 2009).

Visualizing the findings

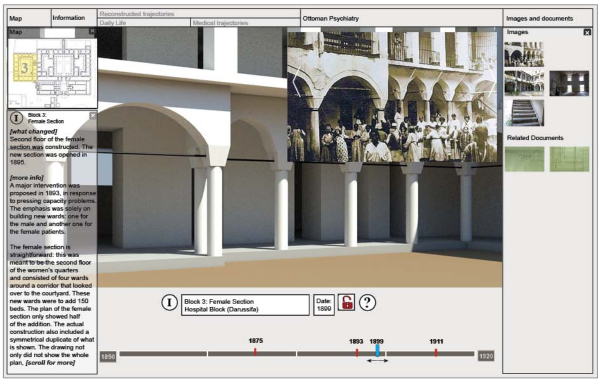
In our project, we faced several challenges to demonstrate the findings with methods traditionally used in architectural history. Ottoman mental facilities in general, and our case study in particular (especially the period that we are focusing on), is largely understudied compared to the aforementioned examples of the Roman Forum or monasteries. All the findings come from primary sources that are incomplete in nature. Moreover, the sources are predominantly textual with limited visual resources describing various states of the building. The historical layers added to the building render conventional visualization methods ineffectual. As a response to these challenges and the resulting complicated presentation of findings, the project employs an interactive 3D simulation toolset built within Unity 3D that allows visualizing hypothetical spatial reconstructions and trajectory tracking all continuously referring to temporal data and incorporating primary documents. Our interactive simulation framework builds on these efforts creating the ability to navigate in four dimensions simultaneously and without restriction:

1 **Temporal dimension:** Instead of the traditional interest in finding out about the "original" building, this approach looks at the **life of a building** and how it changed over time. The method we propose for Unity is to implement a time slider component into the viewing user interface (UI). This component controls the visibility of architectural changes at distinct time periods. In this project, we have identified four specific points in time wherein primary sources identify distinct architectural states. The timeline UI is enabled by including time data within the model data. In the 3D modeling or 3D data acquisition phase, model elements are tagged with custom properties coded to our tool. These custom properties represent abstract metadata defining the start and end date for the object(s) (for temporal properties), the orientation point for the object(s) (for geospatial locating), as well as any descriptive information all of which become native to the file.

2 **Reconstructed trajectories:** Instead of treating space as an isolated entity, this approach captures the **life in the building**. By tracing the movement of the occupiers of the space, we know more about their lives. The everyday life in the asylum consisted of the acts and movements of its occupiers, in this case, patients, doctors and the staff. By tracing their movements and visualizing them, it becomes possible to have a better understanding of the daily routines (eating, cleaning, sleeping etc.) in addition to the medical treatments that took place through a certain reorganization of space and time. This setting is particularly illuminating as it was assumed at the time that the life and routines of patients in the asylum were crucial components of the healing process.The user interface (UI) enables diagrammatic circulation indicators for each of the four points in time to be visible at the viewer's discretion.

3 **Cross-referencing data:** In addition to the ability to preview time and movement, the interactive functionality in Unity allows the viewer to select the information relevant to the simulation viewed at any time. The ability to combine text as hyperlinked visual overlays and graphic information enables important reference texts and images to be associated with spatial and temporal elements and viewed when desired. One advantage of this presentation method is the ability to break free from the linear structure of a textual manuscript. One can interact with the information presented in a nonlinear way (albeit guided). By cross-referencing various types of information, we aim to create a system that binds or connects these documents and data sets in meaningful ways.

We do not intend the spatiotemporal model to be the **primary** outcome, but rather a platform inviting interpretation and scholarship through the combined textual and visual data. In line with the ontological approach adopted in digital humanities, the ever-evolving outcome is a "knowledge representation, a hypothesis" not a recreation of reality (Bonde et al., 2009). The relationship between the representation and its referent also calls into question producing knowledge with partial data, which is not uncommon in (conventional or digital) historical scholarship. Both possibilities and risks exist in 3D visualization with partial data. 3D models give the opportunity to relate to the building without the trained eye to "read" 2D representations. Risk is related: 3D creates a sense of absolute masking the hypothetical elements within. Being aware of this danger, the project aims to visually clarify the "known" versus "unknown" using graphic coloration indicating the extent of "known" detail.



[what changed]
Second floor of the female section was constructed. The new section was opened in 1895.

[more info]
A major intervention was proposed in 1893, in response to pressing capacity problems. The emphasis was solely on building new wards: one for the male and another one for the female patients.

The female section is straightforward: this was meant to be the second floor of the women's quarters and consisted of four wards around a corridor that looked over to the courtyard. These new wards were to add 150 beds. The plan of the female section only showed half of the addition. The actual construction also included a symmetrical duplicate of what is shown. The drawing not only did not show the whole plan, [scroll for more]

Block 3: Female Section
Hospital Block (Darussifa)

Figure 1: Simulation User Interface Mockup (B. Ozludil, U. Thompson, A. Wendell). Photograph source: Sihhat Almanaki, 1933.

The tool we propose visualizes temporally-located spatial data that allows multiple readings and interpretations, and that is open to manipulation, rather than a mere representation of a space. Doing so will open the approach to more disciplines and interested parties, and while still technical in nature, will provide a platform for historical spatial research. This work can be shared with other digital humanities scholars employing Unity or similar tools. Integrating innovative technologies in historical research has the potential to change the ways in which we conceptualize and tackle the problem at hand. In other words, rather than being "tools" to accomplish what is predetermined, they open up new ways to think and to produce knowledge.

Bibliography

Alkhoven, P. (1991). The Reconstruction of the Past: The Application of New Techniques for Visualization and Research in Architectural History. *Computer Aided Architectural Design Futures: Education, Research, Applications*. Zürich (Switzerland), pp. 549-566.

Boeykens, S. and Neuckermans, H. (2009). Architectural design analysis, historical reconstruction and structured archival using 3D models: Techniques, methodology and long term preservation of digital models. In Tidafi, T. and Dorta, T. (Eds.) Joining Languages, *Cultures and Visions: CAADFutures 2009*, PUM, pp. 119-32.

Bonde, S., Clark M., Elli M. and Julia F. (2009). The Virtual Monastery: Re-Presenting Time, Human Movement, and Uncertainty at Saint-Jean-des-Vignes, Soissons. Visual Resources: *An International Journal of Documentation*, **25**(4): 363–77.

Bourdakis, V. and Pentazou, I. (2012). Real City Museum/Virtual City Model: Real Datasets/Virtual Interactions. In Achten, H., Pavlicek, J., Hulin, J. and Matejovska, D. (Eds.), *Digital Physicality - Proceedings of the 30th eCAADe Conference*, **2**, pp. 337-41.

Drucker, J. (2013). Is There a "Digital" Art History?. Visual Resources: *An International Journal of Documentation*, **29** (1-2): 5-13.

Favro, D. and Johanson, C. (2010). Death in Motion: Funeral Processions in the Roman Forum. *Journal of the Society of Architectural Historians*, **69**(1): 12-37.

Gill, A. A. (2009). Digitizing the Past: Charting New Courses in the Modeling of Virtual Landscapes. Visual Resources: *An International Journal of Documentation*, **25**(4): 313-32.

Kalay, Y., Kinayoglu, G. and Kim, S. W. (2005). Spatio-Temporally Navigable Representation and Communication of Urban Cultural Heritage. Proceedings: *VSMM 2005 International Conference on Virtual Systems and Multimedia*. Ghent, Belgium, pp. 145-52.

Osman, M. (1933). Sihhat Almanaki. Kader Matbaasi, Istanbul.

Data Praxis in the Digital Humanities: Use, Production, Access

Thomas George Padilla
tpadilla@msu.edu
Michigan State University, United States of America

Devin Higgins
higgi135@msu.edu
Michigan State University, United States of America

While there have been numerous efforts at framing the history of the Digital Humanities, no study has concretely characterized the extent to which Digital Humanities research is data driven (Gold and Klein, 2012; Schreibman, Siemens, and Unsworth, 2004; Nyhan, Flinn and Welsh, 2015; Terras, Nyhanand Vanhoutte, 2013). Debates related to this topic periodically crop up along the Hack/Yack divide, as recurrent waves of scholars reflect on the varied histories, projects, and positions that comprise the Digital Humanities (Nowiskie, n.d.; Ramsay, n.d.; Cecire, n.d.; Alvarado, 2012). While these debates will likely continue, it is clear that current theoretical and historical contextualization stand to benefit from a more granular evaluation. The benefits of this evaluation hold potential to shed light on data driven research practices across disciplines and academic ranks, distribution of this output by institution type and geographical location, relative research data accessibility, as well as illumination of the scope of data resources utilized to further Digital Humanities research, which in turn holds the potential to inform library efforts to augment Digital Humanities support with more nuanced focus on acquisition, preparation, and provision of data that is more readily usable to Digital Humanists (Bryson et al., 2011; Sustaining the Digital Humanities, n.d.; Rockenbach, 2013). In order to realize these benefits the present study focuses on Digital Humanities praxis that is expressly data driven and computationally contingent. The study of this praxis is achieved through analysis of nearly 500 articles drawn from seven years of Oxford University Press' Digital Scholarship in the Humanities (formerly Literary and Linguistic computing), seven years of Digital Humanities Quarterly (the full run of the journal), as well as the full run of the Journal of Digital Humanities.

In order to evaluate data praxis, it was necessary to come to a working definition of "data" scoped to the level of concrete usage patterns in the Digital Humanities. The conclusion that a particular article utilized source "data" was based on whether or not the material under analysis played a role in supporting research claims predicated on the affordances of the digital object itself. A close reading of a digital version of *Jane Eyre* therefore would not meet the criterion of data driven, but topic modeling *Jane Eyre* would, as this is a form of analysis that is uniquely possible given digital instantiation of the object under study. Articles which were understood more as reports on data-oriented research, rather than active analysis were typically excluded. Assessments, historiographies, and other meta-analysis of computational research represented elsewhere are not treated as data driven for the purposes of this study as the work in question can move forward without leveraging the affordances of a digital object. Even where these types of articles are held to not contain source data under a process of direct computational analysis or representation, they are still considered against a rubric of research data production. Research data is understood to encompass any non-rhetorical, primarily structural data generated as an output that is used to validate research findings (Federal Register Notice Re OMB Circular A-110, n.d.). This might include tabular data, computer code, or survey responses.

Research data production is evaluated in order to come to an understanding of how Digital Humanists provide or do not provide access to generated data they use to support their arguments. The authors sought to focus on this aspect of research given a growing movement by scholars, operating primarily outside of the Humanities, to make their data and code accessible to support reproducibility and transparency (Stodden, Leisch, and Peng, 2014). If research data was produced in a given article, the authors proceeded to evaluate whether or not it was accessible. Research data is only considered to be accessible if the data in question is made available in a format that is machine processable. Therefore, a table of research data or an image of a line graph included in an article as a JPG is not accessible because the format renders the data intractable. Furthermore, a subset of a larger set of data, mainly used to illustrate an aspect of an argument rather than providing access to the unmediated source dataset is held to be inaccessible. Collectively, researchers, librarians, and publishers can use this portion of the study to inform assessment of the extent to which current research and publication practice are in line with how the field aims to articulate the integrity of its research claims.

On the whole, article level analysis is supported by capturing up to 48 descriptive elements for each article in the target corpus. In aggregate this dataset captures the number of data sources used in a given article, the provider of the data, type of provider, data collection name, content type, format, extent, size, publication pre or post 1923, whether research data is produced, whether research data is accessible, the method of access provided if it is accessible, research data URL, research data type, and a range of demographic data that allows disciplinary characterization of data driven practices by scholars and students, the type of institutions they work in, and where in the world they work. Collectively this data will enable the Digital Humanities community to gain a concrete sense of what proportion of Digital Humanities Scholarship as represented in a core set of journals is data driven. This study indicates that current research and publication practice provide insufficient access to research data. Because the evaluation of a data driven article's argument requires access to research data, this scarcity seems especially troubling. Additional pragmatic gains to be had from this study include ready access to all data sources utilized over the past 7 years in core Digital Humanities journals. Ready access to this data holds potential to increase awareness of data for Digital Humanities research and pedagogy in addition to informing library acquisition, preparation, and provision of data that can used to support the Digital Humanities. Through its concrete focus on data praxis, this study provides newly comprehensive insight into data driven practices across the Digital Humanities.

Bibliography

Alvarado, R., (2012). The Digital Humanities Situation, *Debates in the Digital Humanities*,ed. Gold, Matthew and Klein, Lauren, University of Minnesota Press, http://dhdebates.gc.cuny.edu/debates/text/50.

Bryson T. et al. (2011). *Digital Humanities*, SPEC Kit 326, http://publications.arl.org/Digital-Humanities-SPEC-Kit-326/.

Cecire, N.*Works Cited: When DH Was in Vogue; Or, THATCamp Theory*, http://nataliacecire.blogspot.com/2011/10/when-dh-was-in-vogue-or-thatcamp-theory.html. (accessed October 24, 2015).

Federal Register Notice Re OMB Circular A-110, The White House, https://www.whitehouse.gov/node/15587. (accessed October 31, 2015).

Gold, M. K. and Klein, L. (Eds.) (2012). *Debates in the Digital Humanities*, Minneapolis: Univ Of Minnesota Press.

Nowiskie, B.*Eternal September of the Digital Humanities*, http://nowviskie.org/2010/eternal-september-of-the-digital-humanities/. (accessed October 15, 2010).

Nyhan, J., Flinn, A. and Welsh, A. (2015). *Oral History and the Hidden Histories Project: Towards Histories of Computing in the Humanities*, *Digital Scholarship in the Humanities*, **30**(1): 71–85.

Ramsay, S.*On Building*, http://stephenramsay.us/text/2011/01/11/on-building/. (accessed October 24, 2015).

Rockenbach, B. (2013). Introduction, *Journal of Library Administration*, **53**(1): 1–9, doi:10.1080/01930826.2013.756676.

Schreibman, S., Siemens, R. G. and Unsworth, J.(Eds.) (2004). *A Companion to Digital Humanities*, Malden, MA: Blackwell.

Stodden, V., Leisch, F. and Peng, R. D., eds., (2014). *Implementing Reproducible Research*, The R Series, Boca Raton: CRC Press, Taylor and Francis Group.

Sustaining the Digital Humanities, *Ithaka S+R*, http://www.sr.ithaka.org/publications/sustaining-the-digital-humanities/. (accessed October 30, 2015).

Terras, M. M., Nyhan, J. and Vanhoutte, E. (Eds.) (2013). *Defining Digital Humanities: A Reader*, Farnham, Surrey, England : Burlington, VT: Ashgate Publishing Limited ; Ashgate Publishing Company.

Dépasser La Liste : Quand La Bibliothèque Entre Dans La Danse Des Corpus Web

Cynthia Pedroja
cynthia.pedroja@sciencespo.fr
FNSP - Sciences Po, France

Anne L'Hôte
anne.lhote@sciencespo.fr
FNSP - Sciences Po, France

Elise Chapoy
elise.chapoy@sciencespo.fr
FNSP - Sciences Po, France

Elisabeth Levain
elisabeth.levain@sciencespo.fr
FNSP - Sciences Po, France

Depuis 2014, la bibliothèque de Sciences Po développe une offre de service à destination des laboratoires de recherche de l'Institution. Traditionnellement assez éloignée de ces publics, elle essaye d'adapter ses pratiques et de trouver de nouvelles méthodologies de travail pour accompagner les chercheurs dans leurs travaux.

La révolution numérique a introduit un bouleversement dans la constitution des fonds des bibliothèques: comment rendre compte de l'état d'un sujet d'actualité pour les générations à venir lorsque le discours ne se bâtit plus uniquement dans des médias coutumiers, mais dans le flux des réseaux ? La bibliothèque peut-elle opérer des carottages thématiques du web et concevoir les dispositifs qui en conserveraient la trace ?

Les corpus web tentent de répondre à cette double problématique documentaire et de recherche en s'appuyant sur des techniques d'exploration et de visualisation de réseaux web. S'inscrivant dans la lignée des corpus d'étude outillés proposés par Corinne Welger-Barboza, ils permettent de structurer et d'identifier les discours sous-jacents, tout en apportant des moyens d'appropriation adaptés.

De la construction des données aux premiers constats scientifiques, cette communication dresse un bilan de ce travail, expérimental pour la bibliothèque tant au niveau des outils mobilisés et développés que des processus documentaires mis en œuvre, à travers l'exemple d'un corpus des acteurs de la question des changements climatiques.

Construction

En 2015 se tient à Paris la Conférence des Nations Unies sur les changements climatiques, la COP21. Sciences Po a retenu cette année ce sujet d'actualité comme thème fort de l'institution; le contenu des enseignements, des manifestations scientifiques et de vulgarisation, des expositions en est irrigué. La bibliothèque et le médialab[1], partenaires de ce projet, l'ont choisi comme problématique de leur premier corpus. Quels sont les acteurs de la discussion autour des changements climatiques? Quelle est leur position quant à la responsabilité de l'homme? Qui parle avec qui? Faut-il réduire les émissions de gaz à effet de serre? Voici la liste des questions que nous avons posées. Pour y répondre, nous avons développé un processus en 3 temps: construction, exploration et exposition des données.

Le crawler Hyphe est l'outil qui nous permet de constituer les données. Pour cette opération, itérative, deux types d'actions sont mises en œuvre : identifier, sélectionner. La première consiste à trouver des sites (entités) pertinents pour la thématique. C'est le cœur du corpus[2]. Les crawlers parcourent les liens hypertextes de ce cœur afin d'identifier de nouvelles entités. La seconde à distinguer ceux qui sont pertinents et qui seront crawlés, de ceux qui ne le sont pas. De proche en proche, nous prospectons le web et de nouveaux sites apparaissent. De 60 entités de départ, 40.000 ont été identifiées et 600 ont été conservées. Ce sont ces dernières qui forment le corpus (Jacomy, 2015).

Cette étape repose sur des pratiques à la fois connues et nouvelles pour la bibliothèque: l'analogie entre la sélection d'ouvrages à partir de sources fiables (ici pertinentes) est assez aisée à comprendre. En revanche, crawler, et ainsi automatiser l'opération d'identification reste une action assez éloignée des processus traditionnellement mobilisés en bibliothèque. D'autant plus que le crawler utilisé, Hyphe, conserve l'empreinte des liens qui unissent les entités constitutives du corpus: le réseau. Au-delà de la liste des sites de référence sur les changements climatiques, Hyphe nous permet donc de garder la trace des discussions en ligne. La compétence méthodologique sur les réseaux est apportée ici par le médialab.

Exploration

L'exploitation envisagée nécessite que les entités web soient enrichies. Les métadonnées des sites ne sont pas standardisées, à l'inverse des objets traditionnellement traités en bibliothèque (ouvrages, revues, documents audiovisuels) qui sont, eux, décrits avec des langages contrôlés. C'est pourquoi une série de catégories a été créée: type d'acteur, responsabilité de l'homme dans le changement climatique, nature des actions soutenues (atténuation, adaptation).

Ce travail est la combinaison de compétences documentaires et scientifiques. Le chercheur spécialiste de la thématique propose, et valide les catégories choisies, mais il aide également à circonscrire le périmètre du corpus, en optant par exemple pour la conservation de sites anglophones uniquement. Les catégories sont construites sur des index élaborés par la bibliothèque qui les éprouve lors des phases de test. De plus, pour chaque site, un résumé est rédigé afin de garder une trace de son contenu et ainsi pallier la nature éphémère du web (Corey, 2010).

En l'absence d'outil de catégorisation pour ce type de corpus, le développement d'une solution ad hoc est nécessaire. Elle doit répondre à des contraintes claires, dont le travail collaboratif synchrone et le maintien de l'intégrité des données entre Hyphe et le site web.

La phase exploratoire ultime est celle de la visualisation: « faire parler » le réseau en s'appuyant sur la catégorisation[3].

Exposition

Le troisième temps est celui de l'exposition: comment dépasser la liste, par laquelle une bibliothèque présente traditionnellement les ressources web sélectionnées? En proposant un outil d'exploration en ligne, qui permet à l'utilisateur non seulement de consulter les sites du corpus, mais également de se les approprier. En combinant les catégories au graphe, l'usager peut jouer la partition écrite par la bibliothèque: si la liste, rassurante, car connue, est toujours là, il est en plus possible de visualiser, par exemple, les acteurs institutionnels qui soutiennent les mesures d'atténuation des émissions de gaz à effet de serre et de les confronter à celles des blogueurs climato-sceptiques en manipulant la visualisation. Le graphe ajoute de la profondeur à la liste en faisant apparaître la dimension de connexion entre les entités: c'est un atout du corpus.

Les interfaces web existantes exposent rarement cette dualité liste/visualisation. C'est pourquoi, une nouvelle fois, une solution adaptée aux besoins des utilisateurs a été développée[4]. Ce projet entre dans un cadre de libération du code et d'ouverture des données.

Conclusion

Ce travail est le fruit d'une collaboration étroite entre les différents corps de métiers inhérents aux Digital Humanities. 11 personnes ont participé à cette expérimentation : 1 chef de projet, 4 bibliothécaires dont le référent « environnement », 1 développeur, 2 informaticiens spécialistes des réseaux web, 3 chercheurs du domaine. Il aura fallu une année pour mener à terme ce projet qui avait valeur de test: la bibliothèque a-t-elle les moyens de proposer un service d'accompagnement des chercheurs pour l'étude des réseaux web? Répond-elle également à la question de l'archivage du web ?

Le premier constat, positif, est celui de la faisabilité : le collectif a été en mesure de finaliser ce corpus. Le deuxième, tout aussi engageant est celui des premiers résultats scientifiques basés sur la visualisation du réseau. L'équipe de recherche a été surprise de voir que les climato-sceptiques sont encore très actifs sur le web, alors même qu'ils semblent minoritaires dans les discussions physiques. Qui sont-ils? Comment interagissent-ils avec les institutions, les entreprises et la société civile? Pourquoi ne forment-ils pas un groupe à part, mais apparaissent-ils comme parties prenantes de la conversation en ligne? Ces questions doivent maintenant être analysées en profondeur. D'objet documentaire, ce corpus sur le changement climatique est devenu objet d'étude outillé. Chaque carte peut entériner ou infirmer les problématiques initiales et parfois générer des intuitions, interrogations qui doivent être vérifiées et prouvées scientifiquement.

Ce projet a permis à la bibliothèque de collaborer avec un laboratoire de recherche et ainsi d'aller au-delà de ses logiques traditionnelles de documentation. De nouvelles méthodes de travail ont été identifiées. Il s'agit maintenant de les fixer et de voir s'ils sont opérants sur d'autres thématiques en achevant la conception de la chaîne d'outils et en faisant monter en compétence les équipes sur leur utilisation et la connaissance des réseaux web. La phase suivante est également celle de l'archivage: la documentation du processus de construction et d'exploration de chaque corpus est une étape nécessaire pour garder une trace pérenne des ressources web et assurer les conditions de réutilisation des corpus.

S'appuyant sur la décomposition du travail autour de la construction, l'exploration et l'exposition des données, l'équipe a su répondre à la double problématique à laquelle l'institution est confrontée. Et ainsi, comme une valse, la bibliothèque est entrée en trois temps dans la danse des corpus web.

Bibliography

Corey, D. (2014). Archiving the Web: A Case Study from the University of Victoria. *Code4Lib, 26*. http://journal.code4lib.org/articles/10015 (accessed 4 March 2016).

Girard, P. (2011). HyperText Corpus Initiative: how to help researchers sieving the web, *Proceedings of the Out of the Box conference: Using Web Archives Conference*, London, UK. http://spire.sciencespo.fr/hdl:/2441/5coittpe7h8g695h172cg34d3e (accessed 4 March 2016).

Jacomy, M. (2015). L'analyse visuelle de réseaux. Explorer le social grâce au numériqu, *I2D – Information, données and documents*, **52**(2): 60-61. http://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-60.htm (accessed 4 March 2016).

Niu, J. (2012). *Functionalities of Web Archives, D-Lib Magazine*, **18**(3/4). http://dx.doi.org/10.1045/march2012-niu2 (accessed 4 March 2016).

Venturini, T., Baya Laffite, N., Cointet, J., Gray, I., Zabban, V. and De Pryck, K. (2014). Three maps and three misunderstandings: A digital mapping of climate diplomacy, *Big Data and Society*, **1**(2): 1-19. http://spire.sciencespo.fr/hdl:/2441/1lpf2c85nl8a5bvji6lcpcc4bp (accessed 4 March 2016).

Welger-Barboza, C. (2010). Quelques réflexions sur l'effet propédeutique des catalogues des collections des musées en ligne, *DH2010 Conference*, London, UK. https://docs.google.com/viewer?url=http%3A%2F%2Fdh2010.cch.kcl.ac.uk%2Facademic-programme%2Fabstracts%2Fpapers%2Fpdf%2Fbook-final.pdf (accessed 4 March 2016).

Welger-Baroza, C. (n.d.). *Corpus d'étude outillés*. [Blog] Observatoire critique. http://observatoire-critique.hypotheses.org/category/corpus-detude-outilles (accessed 4 March 2016).

Notes

[1] «Laboratoire de recherche centré sur les méthodologies numériques qui présente la particularité de s'appuyer sur une approche théorique forte et originale en sciences sociales, la théorie de l'acteur-réseau» [http://www.medialab.sciences-po.fr/fr/about/]

[2] L'équipe a retenu comme point de départ les 12 thèmes identifiés dans (Venturini, 2014)

[3] Quelques visualisation sont disponibles sur le site http://medialab.github.io/double-dating-data/#/

[4] http://corpusweb.sciencespo.fr/app/#/

RAPSCAPE – un'esplorazione dell'universo linguistico del rap attraverso il text-mining e la data-visualization

Stefano Perna
sperna@unisa.it
Università degli Studi di Salerno, Italy

Raffaele Guarasci
r.guarasci@icloud.com
Università degli Studi di Salerno, Italy

Alessandro Maisto
maisto.ale@gmail.com
Università degli Studi di Salerno, Italy

Pierluigi Vitale
pierluigivitale@hotmail.it
Università degli Studi di Salerno, Italy

In questo lavoro descriviamo un progetto di ricerca, attualmente in corso, dedicato all'analisi dell'universo linguistico e semantico della musica rap, con particolare attenzione rivolta alla realtà italiana. L'obiettivo del lavoro è quello di arrivare ad offrire una mappatura panoramica, una "distant reading" della lingua usata dal rap italiano.

La scelta di questo genere è motivata dal fatto che il rap è tra i fenomeni più vitali e dal maggiore impatto socioculturale della musica e delle sottoculture giovanili degli ultimi decenni (Lena, 1995; Toop, 1999; Forman and Neal, 2004; Pinkney, 2007), esteso ormai ben oltre gli originari confini statunitensi per divenire fenomeno globale (Androutsopoulos and Arno 2003; Osumare, 2007; Alim et al., 2008) e all'interno del quale è possibile riscontrare una ricchissima produzione testuale ed un alto tasso di innovazione e sperimentazione di forme linguistiche (Cutler, 2007; Bradley, 2009; Terkourafi, 2010).

L'idea alla base del lavoro è quella di ottenere una "cartografia" della lingua del rap, che permetta di osservare e analizzare nel suo complesso un settore della produzione culturale contemporanea estremamente diffuso e popolare anche in Italia (Pacoda 1996; Filippone and Papini, 2002; Attolino, 2003; Scholz, 2005). In questo lavoro focalizziamo l'attenzione principalmente sulla dimensione testuale del rap piuttosto che su quella musicale, pur trattandosi di un genere in cui il rapporto tra parola e ritmo è inestricabile (Bradley, 2009). In ogni caso, è possibile affermare che la componente testuale nel rap occupa un ruolo centrale e che la specificità del vocabolario, dei temi, della capacità di invenzione linguistica nonché l'importanza dell'aspetto narrativo (Attolino, 2012) fanno dei testi del rap un corpus linguistico interessante da analizzare in sè.

A tal fine, piuttosto che soffermare l'attenzione su di un numero limitato di testi da analizzare in profondità, mettiamo in campo una metodologia di lavoro multidisciplinare - in cui convergono web data-mining, linguistica e information design – con l'obiettivo di giungere alla costruzione di un database testuale molto ampio da sottoporre ad analisi mediante strumenti di text-mining e di linguistica computazionale e da rendere esplorabile mediante una serie di visualizzazioni interattive elaborate ad-hoc.

In una prima fase si è proceduto all'individuazione di alcune web-repository contenenti le trascrizioni dei testi delle canzoni rap in lingua italiana. Non essendo le fonti ufficiali (siti personali degli artisti, siti delle etichette, libretti dei CD, ecc.) particolarmente ricche di informazioni, sono stati individuati alcuni popolari siti di text-sharing, dove fan e ascoltatori forniscono spontaneamente le proprie trascrizioni dei testi degli artisti.

Sulle fonti selezionate è stato addestrato uno script di web-scraping, sviluppato appositamente, in grado di estrarre, per ogni brano presente sul sito, il testo e i meta-dati di riferimento (titolo brano, nome autore, collaborazioni, album). Una volta addestrato lo script si è passati alla fase di estrazione dati vera e propria che ha portato alla costruzione di un database di circa quindicimila brani. Il risultante database è stato poi sottoposto ad una prima fase di pre-processing e data-wrangling per renderlo disponibile all'analisi successiva. Sul testo estratto dal web è stata effettuata una profonda ripulitura con metodi semi-automatici in modo da ottenere un corpus omogeneo di testi trattabili computazionalmente.

Alla fase di estrazione e standardizzazione del dataset segue la fase di analisi linguistica. In questa fase teniamo conto di alcuni studi precedenti condotti nell'ambito del MIR - Music Information Retrieval, in particolare quelli rivolti all'analisi automatica dei testi delle canzoni (Mahedero et al., 2005; Kleedorfer et al., 2008; Hu et al., 2009) e dei testi rap in particolare (Hirjee and Brown, 2009; Hirjee and Brown, 2010; Malmi et. al, 2015).

Il corpus è processato usando l'intera pipeline di analisi linguistica (Manning and Schütze, 1999) già ampiamente nota nei task di NLP: tokenizzazione, lemmatizzazione e pos tagging. Successivamente si è passato ad un'analisi statistica per ottenere le frequenze assolute dei termini, le frequenze relative per autore, le collocazioni, i bigrammi e i trigrammi ricorrenti e la forza di associazione tra le parole espressa in termini di PMI (Pointwise Mutual Information). Gli strumenti utilizzati per effettuare queste analisi sono basati sulla libreria NLTK in Python (Loper and Bird, 2002). Una volta estratti i Lemmi con le rispettive frequenze, viene calcolato il valore di Term Frequency/ Inverse Document Frequency (TF/IDF) per ogni lemma in modo da estrarre le parole più significative per ciascun autore. Una matrice di co-occorrenza, precedentemente costruita su un corpus di circa 3 milioni di parole, attraverso l'applicazione di un algoritmo di Distributional Semantics chiamato HAL - Hyperspace Analogue to Language (Burges and Lund, 1995), è utilizzata per estrarre le parole con valori di similitudine semantica maggiori per ogni lemma, allo scopo di creare un network di significati che identifichi lo spazio semantico di ciascun autore e permetta la loro classificazione attraverso algoritmi di machine learning (clustering).

I dati risultanti dall'analisi linguistica sono strutturati in un database adatto all'elaborazione dei software e dei processi di data visualization. L'obiettivo di questa parte del progetto è quello di costruire un tool interattivo che utilizzi tecnologie web (html, css, javascript) per rendere il dataset esplorabile, comunicabile e analizzabile ulteriormente. Per l'elaborazione del sistema di visualizzazione prendiamo in esame le specifiche problematiche poste dalla visualizzazione di grandi corpora testuali (Wise et al., 1995; Fortuna et al., 2005; Alencar et al., 2012; Sinclair et al., 2013; Kucher, 2014; Brath and Banissi, 2015) e le soluzioni approntate da alcuni lavori precedenti sulla visualizzazione di database composti da testi di canzoni (Labrecque, 2009; Baur et al., 2010; Oh, 2010; Sasaki et al., 2014).

Il tool di visualizzazione si compone di una serie di "viste" e di filtri di navigazione che permettono di osservare il dataset da più angolazioni e attraverso diversi livelli di dettaglio, secondo il classico pattern Overview first, zoom and filter, then details-on-demand (Shneidermann, 1996). Oltre agli approcci classici dell'Information Visualization, la progettazione delle visualizzazioni tiene conto dell'approccio maturato dal design della comunicazione nell'ambito delle Digital Humanities (Uboldi and Caviglia, 2015) in paritcolare per quanto riguarda la definizione della user experience.

Una serie di layer di visualizzazione sono combinati in delle viste panoramiche che offrono uno sguardo complessivo su diversi aspetti del database: statistiche di base come le frequenze e le distribuzioni dei termini più utilizzati; la varietà complessiva del vocabolario; una serie di ranking; reti bipartite tra autori e termini; reti tra parole e relativi cluster semantici più evidenti.

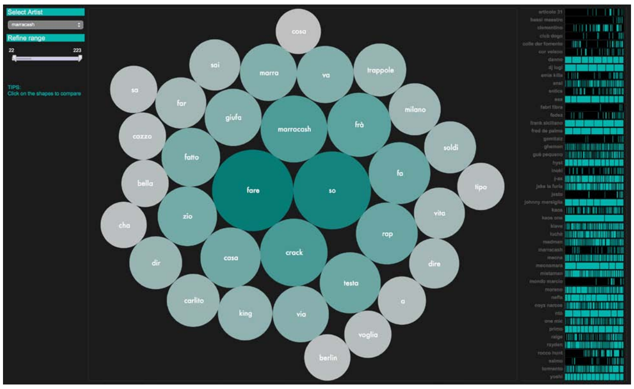


Fig : Rapscape: Single Artist Explorer view

Filtri e viste secondarie sono progettati invece per muoversi rapidamente tra diversi livelli e prospettive sul dataset e di scendere nel dettaglio per analizzare i dati relativi al singolo autore (termini più frequenti, temi dominanti, ecc) o al singolo brano. E' inoltre possibile operare comparazioni tra autori (o gruppi di autori), o tra brani (o gruppi di brani). La visualizzazione è progettata dunque principalmente come strumento esplorativo in modo tale da rendere possibile l'analisi dell'universo testuale del rap a diversi livelli di profondità e granularità.

Bibliography

Alencar, A. B., Oliveira, M. C. F. de and Paulovich, F. V. (2012). Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**(6): 476–92.

Alim, H. S., Ibrahim, A. and Pennycook, A. (2008). *Global Linguistic Flows: Hip Hop Cultures, Youth Identities, and the Politics of Language*. Routledge.

Androutsopoulos, J. and Scholz, A.(2003). Spaghetti Funk: Appropriations of Hip-Hop Culture and Rap Music in Europe. *Popular Music and Society*, **26**(4): 463–79.

Attolino, P. (2003). *Stile ostile*, CUEN.

Attolino, P. (2012). Iconicity in Rap Music The challenge of an anti-language. *Testi e Linguaggi*, **6**: 17–35.

Baur, D., Steinmayr, B. and Butz, A. (2010). SongWords: Exploring Music Collections Through Lyrics. In *ISMIR*, pp. 531-36.

Bradley, A. (2009). *Book of Rhymes: The Poetics of Hip Hop*. Basic Books.

Brath, R., and Banissi, E. (2015). Using Text in Visualizations for Micro / Macro Readings. *Proceedings of the IUI Workshop on Visual Text Analytics*. Retrieved from http://vialab.science.uoit.ca/textvis2015/papers/Brath-textvis2015.pdf

Burgess, C., and Lund, K. (1995). Hyperspace analog to language (hal): A general model of semantic representation. In *Proceedings of the annual meeting of the Psychonomic Society*, **12**: 177-210.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.*, **16**(1): 22–29.

Cutler, C. (2007). Hip-Hop Language in Sociolinguistics and Beyond. *Language and Linguistics Compass*, **1**(5): 519–38.

Filippone, A. and Papini, L. (2002). La parola e il suo potere: Il linguaggio del rap italiano. *Rassegna italiana di linguistica applicata*, **33**(3): 71–86.

Forman, M. and Neal, M. A. (2004). *That's the Joint!: The Hip-Hop Studies Reader*. Psychology Press.

Fortuna, B., Grobelnik, M. and Mladenić, D. (2005). *Visualization of text document corpus.Informatica*, pp. 497–502.

Hirjee, H. and Brown, D. G. (2009). Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pp. 711–16.

Hirjee, H. and Brown, D. G.(2010). Rhyme Analyzer: An Analysis Tool for Rap Lyrics. *Proceedings of the 11th International Society for Music Information Retrieval Conference*.

Hu, X., Stephen, J., Andreas, D. and Ehmann, F. (2009). Lyric text mining in music mood classification. *Proceedings of the International Society for Music Information Retrieval Conference*.

Kleedorfer, F., Knees, P. and Pohle, T. (2008). Oh Oh Oh Whoah! Towards Automatic Topic Detection In Song Lyrics. *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pp. 287–92.

Kucher, K., and Kerren, A. (2014). Text Visualization Browser: A Visual Survey of Text Visualization Techniques. In *IEEE Information Visualization (InfoVis' 14), Paris, France.*

Labrecque, A. (2009). *Computer Visualization of Song Lyrics*. Doctoral dissertation: Worcester Polytechnic Institute.

Lena, J. C. (2006). Social Context and Musical Content of Rap Music, 1979–1995. *Social Forces*, **85**(1): 479–95.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. (ETMTNLP '02). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 63–70

Mahedero, J. P. G., MartÍnez, Á., Cano, P., Koppenberger, M. and Gouyon, F. (2005). Natural Language Processing of Lyrics. *Proceedings of the 13th Annual ACM International Conference on Multimedia*. (MULTIMEDIA '05). New York, NY, USA: ACM, pp. 475–78.

Malmi, E., Takala, P., Toivonen, H., Raiko, T. and Gionis, A. (2015). DopeLearning: A Computational Approach to Rap Lyrics Generation. *arXiv:1505.04771 [cs]* http://arxiv.org/abs/1505.04771 (accessed 5 March 2016).

Manning, C. D., and Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT press.

Oh, J.(2010). Text Visualization of Song Lyrics. *Center for Computer Research in Music and Acoustics, Stanford University*.

Osumare, H. (2008). *The Africanist Aesthetic in Global Hip-Hop: Power Moves*. Palgrave Macmillan US.

Pacoda, P. (1996). *Potere alla parola: antologia del rap italiano*. Feltrinelli.

Pinckney, C. (2007). *The Influence of Hip-Hop Culture on the Perceptions, Attitudes, Values, and Lifestyles of African-American College Students*. ProQuest.

Sasaki, S., Yoshii, K., Nakano, T., Goto, M., and Morishima, S.(2014). LyricsRadar: A Lyrics Retrieval System Based on Latent Topics of Lyrics. In *ISMIR*, pp. 585-90.

Scholz, A.(2005). *Subcultura e lingua giovanile in Italia: hip-hop e dintorni*. Aracne.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. *IEEE Symposium on Visual Languages, 1996. Proceedings*, pp. 336–43.

Sinclair, S., Ruecker, S., and Radzikowska, M.(2013). Information Visualization for Humanities Scholars. *Literary Studies In The Digital Age*.

Terkourafi, M. (2010). *The Languages of Global Hip Hop*. A&C Black.

Toop, D. (2000). *Rap Attack 3: African Rap to Global Hip Hop*. Serpent's Tail.

Uboldi, G. and Caviglia, G. (2015). Information Visualizations and Interfaces in the Humanities. In Bihanic, D. (Ed), *New Challenges for Data Design*. Springer London, pp. 207–18 http://link.springer.com/chapter/10.1007/978-1-4471-6596-5_11 (accessed 5 March 2016).

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. *Information Visualization, 1995. Proceedings.* pp. 51–58.

Metaphor Mining in Historical German Novels: Using Unsupervised Learning to Uncover Conceptual Systems in Literature

Stefan Pernes
stefan.pernes@uni-wuerzburg.de
University of Würzburg, Germany

Figurative language poses a challenge to Natural Language Processing systems, while being a ubiquitous phenomenon that is deeply ingrained in every-day language. As corpus studies suggest, figurative language devices appear on average in every third sentence of general-domain text (Shutova, 2015), thus making the development of automatic recognition and interpretation systems play an important role in many text mining use cases, especially those aiming for a deeper semantic understanding of texts. Furthermore, acknowledging the pervasiveness of such language forms - and of the emblematic device of metaphor in particular - allows for a change in perspective, not conceiving it as a merely rhetorical device but as a genuinely cognitive mechanism that manifests itself in language in the form of surface metaphorical expressions. Such surface expressions usually follow a directionality principle common in figurative language which is to project one domain of experience (the source, e.g. war) onto another (the target, e.g. argument), the first one typically being more concrete and the second one being more abstract. Taken together, such surface expressions (e.g. She shot down all of my arguments) consitute a conceptual metaphor argument is war, a cognitive phenomenon that can be studied through its expression in language. This line of thought has been established as Conceptual Metaphor Theory (Lakoff and Johnson, 1980), by now a widely adopted and empirically grounded approach (e.g. Gibbs, 2008) that has opened up a interdisciplinary field of research, not least with involvement from computational linguistics.

Analysing metaphorical language use from a (cognitive) anthropologic and psycholinguistic point of view has various possible applications: The approach qualifies for research questions from the fields of critical discourse analysis, media studies, and philosophy, as it sheds light on a collective subconscious, encompassing ideological subtexts, and maybe even pre-discursive existential territories (Guattari, 2008) as traced out in late 20th century philosophy. Another area of application is text classification in literary studies: Found metaphorical expressions and conceptual mappings can be used as features to describe the relative similarity of observed texts and thus lend themselves to genre identification and authorship attribution (Lodge, 1988).

The research described here takes up this theoretical framework and builds upon a computational metaphor identification and aggregation approach as proposed by Shutova and Sun (2013). Unsupervised machine learning, namely a hierarchical soft clustering strategy known as Hierarchical Graph Factorization Clustering (HGFC), is employed to build up a graph of concepts that reflects aggregate metaphorical mappings. Using conceptual metaphor as a unit of observation allows for a sensible aggregation and tracing of surface metaphorical expressions in large scale corpora, and in this case is also used to follow diachronic developments in a corpus of historical German literature. Furthermore, as a correlate of cognitive processes it should provide an empirically grounded access to the conceptual systems, e.g. cultural and moral models, of examined texts and their times.

The main idea of the approach is to cluster nouns - which are taken to be concepts - according to their selectional preferences, that is, "the tendency for a word to semantically select or constrain which other words may appear in a direct syntactic relation with it" (Roberts and Egg, 2014). In the resulting clustering, figurative language use becomes visible as violation of the most frequent selectional preferences representing the normal, non-figurative case. It is an approach that determines the metaphorical in relation to the normal, which also entails that a sufficient amount of non-metaphorical language use needs to be present in the data. In the case of a diachronic corpus of literature that means to balance the corpus using historical dictionaries and encyclopaedias in order to introduce more prosaic language use.

The dataset is drawn from a large text collection (The Digital Library, 2016) and contains up to 1700 German novels from the early 16th up to the beginning of the 20th century. Preprocessing consists of POS-tagging, lemmatization, and dependency parsing, allowing for an extraction of nouns and their corresponding verbs according to certain grammatical relations - subject, direct object, and indirect object relations. Verbal constructions are only one type of realization, but they do cover a significant part of metaphors usually encountered in the wild. Furthermore, it should be straightforward to generalize the approach in order to include adjectival constructions and similes, which would allow to cover most of the possible metaphorical expressions. Preprocessing is performed using a modular pipeline (Jannidis et al., forthcoming), tailored to the processing of book-length German texts. Subsequently, a number of most frequent nouns (e.g. 2000) and corresponding verbs are extracted. The verbs act as features for the concept clustering and can come from various sources, not necessarily the same corpus as the most frequent nouns. This could be used as a way to introduce balancing text types into the model, without altering the concept graph as derived from the literary corpus. The resulting noun-verb feature matrix is then normalized for each noun vector to sum to 1 and the Jensen-Shannon divergence between pairs of noun vectors is used as a measure to calculate the similarity matrix (the initial concept graph).

With the similarity matrix in place, clustering methods can be applied in order to generate a suitable tree of concepts. Different approaches were tested at this point (using implementations from Python machine learning library scikit-learn, cf. Pedregosa et al., 2011): 1) connectivity-based or agglomerative clustering, which includes average, complete, and - the baseline from Shutova and Sun (2013) - Ward linkage 2) density-based clustering, namely DBSCAN and HDBSCAN, and 3) for subspace-based methods, spectral clustering, as well as spectral bi- and co-clustering. Results where manually reviewed and an internal evaluation measure, the silhouette coefficient, was used to assess the quality of generated clusterings. Results indicate that in this setting, spectral clustering performs very similar to the baseline, while the other methods produce clusterings of inferior quality. This exploration of readily available methods shed some light onto the requirements for unsupervised metaphor identification and aggregation. In addition, tests with balancing and pruning were performed on smaller development corpora: Solely using encyclopedias produces a model that contains mostly synonym and antonym relations but no metaphorical mappings. Similarly, models consisting only of literary texts can lack non-figurative uses for concepts. What can also be observed is that the balancing leads to deeper models, e.g. concepts accumulate more features and aggregate better.

To give an intuition, example clusters from the baseline results on a subset of 383 novels are reproduced here, showing the top ten features for each concept:

IDEAS ARE FOOD

education / bildung (10): geben-dobj beanspruchen-dobj taxieren-dobj voraneilen-subj überstrahlt-dobj ausspräch-subj nahestehen-subj heraustreiben-dobj ermangelnd-dobj abschöpfen-dobj

memory / erinnerung (48): geben-dobj wachzurufen-dobj stören-dobj mahnen-dobj aufgrischen-dobj verlöschen-subj wiederzuerwecken-dobj neubeleben-dobj frischen-dobj hervorschießen-subj

hunger / hunger (10): geben-dobj erweren-dobj büssen-dobj schaben-subj überhen-iobj verschmachten-dobj stärkern-dobj bittern-subj hinausgetreiben-subj trainieren-iobj

EMOTIONS ARE PLANTS

flower / blume (47): pflücken-dobj liegen-subj lieben-dobj begießen-dobj welken-dobj duften-dobj duftet-subj durchhauten-subj hingesenken-subj erblüht-dobj

emotion / gefühl (90): liegen-subj ersticken-dobj abstumpfen-dobj hervorraufen-dobj halten-iobj entspinnen-dobj hinausdehnen-dobj aufwekken-dobj anhielen-subj arten-subj

In principle, cluster labels are manually assigned using categories from Lakoff's master metaphor list (Lakoff et al., 1991). Such is the case with the first example, IDEAS ARE FOOD, while the second one, EMOTIONS ARE PLANTS, is not present in the list and was created to appropriately describe the cluster.

Pending work includes testing HGFC and providing means to include metadata for modeling the diachronicity of the data. Considering the time span covered by the corpus, some orthographic and lexical variation will have to be handled, either by use of a specialized spelling normalization system or a more rigoros treatment such as stemming. It can be noted that HGFC combines some of the characteristics exhibited by the surveyed approaches and running it on the full size corpus will significantly improve on the baseline in terms of the amount of metaphorical expressions and conceptual mappings induced. The system will be evaluated using either a small gold standard of annotated sample sentences or manually compiled conceptual mappings in a confined domain (e.g. using Lakoff et al., 1991), which should give some indication of its precision in the domain of historical German literary texts.

Bibliography

Gibbs, R. W. (2008). Metaphor and Thought. The State of the Art. In Gibbs, R.W. (Ed), *The Cambridge Handbook of Metaphor and Thought.* Cambridge University Press, pp. 3-14.

Guattari, F. (2008). *The Three Ecologies*. Continuum.

Jannidis, F., Reimers, N., Vitt, T., Pernes, S. and Pielström, S. (forthcoming). DARIAH-DKPro-Wrapper Output Format (DOF) Specification. *DARIAH-DE Working Papers*.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live by*. University of Chicago Press.

Lakoff, G., Espenson, J. and Schwartz, A. (1991). *The Master Metaphor List*. University of California at Berkeley.

Lodge, D. (1988). *The Modes of Modern Writing: Metaphor, Metonymy, and the Typology of Modern Literature*. University of Chicago Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, **12**: 2825-30.

Roberts, W. and Egg, M. (2014). A Comparison of Selectional

Preference Models for Automatic Verb Classification. *Proceedings of EMNLP 2014*, pp. 511-22.

Shutova, E. (2015). Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, **41**(4): 579-623.

Shutova, E. and Sun, L. (2013). Unsupervised Metaphor Identification Using Hierarchical Graph Factorization Clustering. *Proceedings of NAACL-HLT 2013*, pp. 978–88.

The Digital Library (2016). *The Digital Library in TextGrid.* https://textgrid.de/en/digitale-bibliothek

When Traditional Ontologies are not Enough: Modelling and Visualizing Dynamic Ontologies in Semantic-Based Access to Texts

Silvia Piccini
silvia.piccini@ilc.cnr.it
ILC-CNR, Italy

Matteo Abrate
matteo.abrate@iit.cnr.it
IIT-CNR, Italy

Clara Bacciu
clara.bacciu@iit.cnr.it
IIT-CNR, Italy

Andrea Bellandi
andrea.bellandi@ilc.cnr.it
ILC-CNR, Italy

Emiliano Giovannetti
emiliano.giovannetti@ilc.cnr.it
ILC-CNR, Italy

Lorenzo Mancini
lorenzo.mancini@ilc.cnr.it
ILC-CNR, Italy

Andrea Marchetti
andrea.marchetti@iit.cnr.it
IIT-CNR, Italy

Introduction

The work described in this paper came about as a result of reflections made within the “Clavius on the web” Project[1], which studied the correspondence between the Jesuit mathematician and also astronomer and some important scientists of his century, such as Galileo and Brahe[2]. One of the main aims of the project is to make it possible for students and scholars to access the texts on a semantic basis, in order to allow a deeper understanding of the often complex content, they convey.

Texts are often the unique source that scholars have at their disposal in order to be able to reconstruct and more completely understand the past author’s thought.

In order for technology to come to the aid of scholars in this effort, the concepts evoked within the text, as well as the terms representing these concepts need to 1) have a structured organization 2) be explicitly and univocally represented and 3) be defined through the relationships that unite them. In order to achieve this, we chose to adopt an ontologybased model, as ontologies are a de facto standard for knowledge representation.

Interestingly, the choice to use ontologies raised some issues, also with regard to theoretical aspects: indeed, standard ontological formalisms usually static and crisp proved to be inadequate in modelling the complexity of the knowledge conveyed by the analysed texts. As a result, more refined models as well as appropriate graphical representations needed to be introduced so that computers would be able to process these ontologies and visualize them in a way that students and scholars could understand and work with them.

The ontological model

Here we list and briefly describe the main aspects of the knowledge conveyed by the Clavius’ corpus that our ontological model should capture.

- **Explicit** versus **implicit knowledge** : our ontology is designed to structure both the entities explicitly evoked in the text (typically denoted by terms) and the entities implicitly entailed as belonging to the background knowledge that the writer implies (which can be possessed by the reader themselves only in part).
- **Shared** versus **individual knowledge** : different authors can share, and in fact do share, some aspects of conceptualising the domain, as clearly they have certain theories and beliefs in common. However, our ontology must formally structure the author’s own conceptualisation of the world, as it emerges from specific textual passages of the analysed corpus.
- **Certain** versus **uncertain knowledge** : in the case where the authors express confidence in some theories or reject and advance doubts towards others. It is therefore essential for each entity which populates the ontology (a class, an instance, a property) to be associated with a degree of certainty.
- **Static** versus **dynamic knowledge** : correspondence implies sharing information and knowledge, which can lead to changes in the way the correspondents view the world, sometimes significant. This is particularly the case

with scholars. As a result, the ontology needs to be dynamic and temporal, so that it is possible to illustrate the evolution of the author's conceptualization over a period of time. The specific time is either explicitly indicated by the author in his/her work or reconstructed from other sources.

Other parameters could be considered, such as vagueness, ambiguity and sincerity. The validity of these aspects is not limited to these kinds of texts (i.e. scientific letters), but it applies to any text such as essays, scientific journals, diaries, which expresses an author's firmly-held or evolving opinion. Consequently, as a case study (see Section 4), we chose Galileo's Sidereus Nuncius(Galilei, 2001), to prove the applicability of the model outside the epistolary corpus. In the present paper, we will mainly focus on dynamic knowledge and its representation.

Models for representing dynamic knowledge

In literature, the problem of representing dynamically evolving information in ontologies has been addressed by adopting several different approaches (Flouris et al., 2008). A very simple solution is to create a version of the ontology for each temporal event that has to be represented (ontology versioning). However, a versioning algorithm is necessary in order to access the different temporal variants of the ontology. Other proposals aim to extend OWL ontologies in order to provide binary relation instances with a time reference. Related approaches are: (Welty et al., 2006) encoding a perdurantist/4D view in OWL, (Krieger, 2008) interpreting original entities as time slices, and (Manola et al., 2004) reifying original relations. For an exhaustive list of works, see (Krieger and Declerck, 2015). However, all of these approaches typically invalidate standard OWL reasoning, and they do not allow the representation of the change in subsumption and instantiation. In (Rizzolo et al., 2009) time semantics is added also to resources by providing temporalvarying classes and individuals, but only for RDF(S) ontologies, by extending the model presented in (Gutierrez et al., 2005). However, domain expertoriented tools for manipulating RDF(S) do not currently exist.

Against this background, we chose to conduct our first experiments with a reification-based approach and SKOS[3], the latter providing the best compromise between temporal aspects representation, availability of tools, querying and reasoning capabilities.

A case study

We propose here a possible representation of the evolution of Galileo's conceptualization of Jupiter's moons over a specific week in 1610, reconstructed on the basis of the Sidereus Nuncius.

The first observation of the planets dates back to 7th January 1610, when Galileo first saw what he thought were three fixed stars near Jupiter. After several observations on 11 th January, he noticed that their position relative to Jupiter changed in the same way as wandering stars. Two days later, he observed that there existed four satellites orbiting around Jupiter and not three.

Here we present the preliminary version of the ontology which structures the content of portions of the Sidereus Nuncius where Galileo describes his observation of Jupiter's moons.

We first identified the key terms of the text as the terminology (in bold in Fig. 2) upon which we defined the explicit entities of the ontology. In addition we specified the necessary implicit entities to add to it (eg. *Galilean moon*). The ontology was built using Protégé 5.0.0 (Musen et al., 2000) and the plugins Skos Editor[4] and Chronos (Preventis et al., 2014), the former to implement an SKOS ontology and the latter to add the diachronic component. The process is described in the following steps:

1. Structuring of the concepts via the skos: broader relation; the concept *Galilean_Moon* has been set as a subconcept of both *Fixed_Star* and *Wandering_Star* (Fig. 1.a);

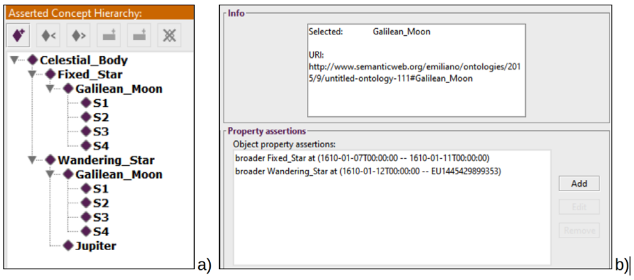
2. Definition of the properties *isNearTo* and *revolvesAround;*

3. Instantiation of these two properties between the four moons (*S1*, *S2*, *S3*, and *S4*) and *Jupiter;*

4. Conversion of the properties broader, *isNearTo* and *revolvesAround* into temporal;

5. Attribution of the correct time interval to each property instantiation.

As a result of this representation, the concept *Galilean_Moon* became narrower than *Fixed_Star* during the time interval between 7th and 11th January 1610, then it changed to narrower than *Wandering_Star* (Fig. 1.b). Analogously, each of the three moons progressed from being simply "nearTo" *Jupiter* to "revolvesAround" *Jupiter*. Finally, starting from 13th January, the relation broader also links *S4* and *Galilean_Moon* (i.e. Galileo spots a fourth object).



Asserted Concept Hierarchy:

- Celestial_Body
  - Fixed_Star
    - Galilean_Moon
      - S1
      - S2
      - S3
      - S4
  - Wandering_Star
    - Galilean_Moon
      - S1
      - S2
      - S3
      - S4
    - Jupiter

Info

Selected: Galilean_Moon

URI:
http://www.semanticweb.org/emiliano/ontologies/2015/9/untitled-ontology-111#Galilean_Moon

Property assertions

Object property assertions:

broader Fixed_Star at (1610-01-07T00:00:00 -- 1610-01-11T00:00:00)
broader Wandering_Star at (1610-01-12T00:00:00 -- EU1445429899353)

Fig. 1: a) The concept hierarchy shown in the "SKOS view" tab of Protégé; b) the temporalized relation "broader" applied to Galilean_Moon.

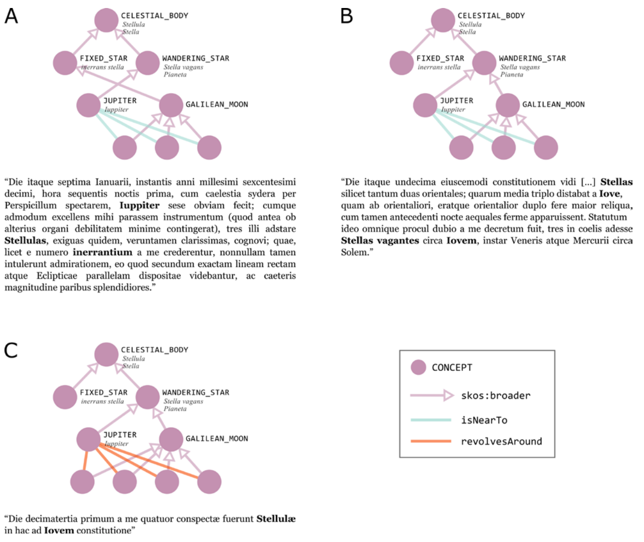
Browsing the constructed dynamic ontology allows to answer to complex queries such as: "how did Galileo's vision of Jupiter's moons evolve in time?" or "which had been Galileo's main changes of perspective about Jupiter in January of 1610?".

Visualization of the ontology

A visualization can be described as an artefact that helps humans to make decisions, learn and communicate, acting as a visual cognitive support (Card et al., 1999). A visual representation of ontologies can therefore be developed to ease their comprehension by both scholars and non-expert users. In our case, a suitable graphical display allows to make visual comparisons between the different time frames of a dynamic ontology, capturing the evolution of the author's ideas. Among the available visualization techniques, we adopted the node-link diagram, which is particularly well suited for exploring the topology of a network and for locating paths (Munzner, 2014).

We wanted to automatically produce visualizations with a quality resembling that of hand-made diagrams (Dwyer et al., 2006; Kieffer et al., 2013). To do so, we observed the work of experts sketching some ontologies on paper, and derived a series of geometric constraints for an automatic placement algorithm. Since the skos:broader relationship defines a *quasi-hierarchy*, i.e., a tree with a reasonably small amount of nodes having multiple parents, the constraints we implemented were intended to produce a familiar, tree-like representation.

The input for the layout algorithm is the entire SKOS graph, and the output is a single layout for all the time frames. Comparison is then made possible by displaying a series of juxtaposed views, each showing only the items of a specific time frame (Fig. 2). This technique ensures that the same item is given the same position in each view, while differences create easy-to-spot “holes”', thus leveraging the user's spatial memory to carry out the comparison task (Munzner, 2014).



“Die itaque septima Ianuarii, instantis anni millesimi sexcentesimi decimi, hora sequentis noctis prima, cum caelestia sydera per Perspicillum spectarem, **Iuppiter** sese obviam fecit; cumque admodum excellens mihi parassem instrumentum (quod antea ob alterius organi debilitatem minime contingerat), tres illi adstare **Stellulas**, exiguas quidem, veruntamen clarissimas, cognovi; quae, licet e numero **inerrantium** a me crederentur, nonnullam tamen intulerunt admirationem, eo quod secundum exactam lineam rectam atque Eclipticae parallelam dispositae videbantur, ac caeteris magnitudine paribus splendidiores.”

“Die itaque undecima eiuscemodi constitutionem vidi [...] **Stellas** silicet tantum duas orientales; quarum media triplo distabat a **Iove**, quam ab orientaliori, eratque orientalior duplo fere maior reliqua, cum tamen antecedenti nocte aequales ferme apparuissent. Statutum ideo omnique procul dubio a me decretum fuit, tres in coelis adesse **Stellas vagantes** circa **Iovem**, instar Veneris atque Mercurii circa Solem.”

“Die decimatertia primum a me quatuor conspectæ fuerunt **Stellulæ** in hac ad **Iovem** constitutione”

Fig. 2: A prototype visualization of the case study presented in section 4. Scholars or students can see the evolution of Galileo's concepts after each observation of Jupiter's moons. The automatically computed diagram layout ensures comparability while preserving a familiar, tree-like appearance.

Bibliography

Card, S. K., Mackinlay, J. D. and Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann.

Dwyer, T., Koren, Y. and Marriott, K. (2006). IPSep-CoLa: An incremental procedure for separation constraint layout of graphs, *Visualization and Computer Graphics, IEEE Transactions on*, **12** (5):821–28.

Flouris, G. et al. (2008). Ontology change: Classification and survey, *The Knowledge Engineering Review*, **23**(2): 117–52.

Galilei, G. (2001). *Sidereus nuncius*, Andra Battistini., Marsilio, Venezia.

Gutierrez, C., Hurtado, C. and Vaisman, A. (2005), Temporal rdf *The Semantic Web: Research and Applications*, Springer, pp. 93–107.

Kieffer, S. et al. (2013), Incremental grid-like layout using soft and hard constraints, *Graph Drawing*, Springer, pp. 448–59.

Krieger, H.U. (2008). Where temporal description logics fail: Representing temporally-changing relationships, *KI 2008: Advances in Artificial Intelligence*, Springer, pp. 249–57.

Krieger, H.-U. and Declerck, T. (2015). An OWL Ontology for Biographical Knowledge. Representing Time-Dependent Factual Knowledge, *Proceedings of the First Conference on Biographical Data in a Digital World 2015*, CEURS-WS.org.

Manola, F., Miller, E. and McBride, B. (2004). RDF primer, *W3C Recommendation*, **10**(1-107): 6.

Munzner, T. (2014). *Visualization Analysis and Design*, CRC Press.

Musen, M.A. et al. (2000). Component-based support for building knowledge-acquisition systems, *Conference on Intelligent Information Processing (IIP 2000) of the International Federation for Information Processing World Computer Congress (WCC 2000)*, **194**.

Preventis, A., Petrakis, E.G. and Batsakis, S. (2014), CHRONOS Ed: a tool for handling temporal ontologies in protégé, *International Journal on Artificial Intelligence Tools*, **23**(4): 1460018.

Rizzolo, F. et al. (2009). Modeling concept evolution: a historical perspective, *Conceptual Modeling-ER 2009*, Springer, pp. 331–45.

Welty, C., Fikes, R. and Makarios, S. (2006). A reusable ontology for fluents in OWL, *FOIS*, **150**: 226–36.

Notes

[1] http://claviusontheweb.it

[2] Clavius' correspondence is contained in the manuscripts APUG 529-530, preserved in the Historical Archives of the Pontifical Gregorian University.

[3] Simple Knowledge Organization System - http://www.w3.org/2004/02/skos/

[4] https://code.google.com/p/skoseditor/

Picture to Score: Driving Vector Animations with Music in the XML Ecosystem

Stephen Ramsay
sramsay.unl@gmail.com
University of Nebraska-Lincoln, United States of America

Brian Pytlik-Zillig
bzillig1@unl.edu
University of Nebraska-Lincoln, United States of America

Audio can be united with video using a number of different techniques. Among the most common are "score to picture" and procedural generation.

"Score to picture" is a feature of most modern DAWs (Digital Audio Workstations), such as *Pro Tools, Logic, Cubase,* and *REAPER*. A composer plays forward the video—usually in the very advanced stages of post production—and sets cues within the software around which a musical soundtrack can be structured. Thus a composer might set a cue to indicate suspense leading up to a particular moment, or the beginning and end of a romantic scene that should be accompanied with incidental music.

Procedural generation goes the other way. Here, a composer creates music—often in a sophisticated audio synthesis environment like *Max/MSP, Pd, Impromptu,* or *SuperCollider*—and uses properties of the audio signal or of the overall program flow to cue events in a video presentation. Since these are full-fledged (if visual) programming languages, driving video with them often means combining the complexity of software engineering with the complexity of handling audio and video signals.

In this short presentation, we describe our experiments with a method of uniting audio and video that lies somewhere between these two approaches. Unlike the practice associated with contemporary filmmaking, our method begins with a musical score and uses events indicated within it as the set of cues for an animation. Rather than use procedural programming or digital signal processing to inform the creation of cues, we use the ordinary conventions of Western musical notation.

To accomplish this, we first represent the score in MusicXML. This might seem an odd choice, given that the MIDI (Musical Instrument Digital Interface) standard was designed precisely to indicate performance events over time. MusicXML, by contrast, was primarily conceived as a way of providing interoperability among software for rendering musical scores as printable objects. Yet MusicXML contains, as one explanation of the standard puts it, a "MIDI-compatible part" concerned with how the music should sound (as opposed to how it should look) (MakeMusic, 2016).

Our system exploits these MIDI-compatible elements—along with several other features of the markup—in order to indicate where a change might occur in an animation. In this way, we are able to use such things as rehearsal marks (sectional markings intended to make it easier for conductors to refer to particular passages), tempo markings, indications of changes in volume (amplitude), emphases, articulations, orchestration, and other aspects of musical notation as cues. And since everything about the duration of a piece and the relationship of the cues within the piece are discernable from the MusicXML file alone, we are able to produce SVG animations that are perfectly in sync with the music from which they are "generated". In the simplest case, this might involve simple changes in color or the movement of shapes, but the system is fully capable of quite advanced 2D animation.

From an artistic standpoint, our way of doing things hearkens back to the earliest days of animation when popular short films were synced to the music of Wagner, Rossini, and Dukas. In this sense, ours is perhaps a new way of doing an old-fashioned thing. But unlike earlier eras, artists today have access to very sophisticated tools for producing digital art. Digital artists regularly use vector graphics programs (like Adobe *Illustrator* and the free *Inkscape)* that can generate SVG, and scoring programs (like *Finale, Sibelius,* and the free *MuseScore)* that can generate MusicXML. What is missing, we think, is a robust way to bridge these two technologies.

Our system provides a very sophisticated bridge in the form of *Indigo*—an SVG animation system developed at CDRH that we have recently re-engineered along the lines we illustrate above. In this presentation, we briefly explain how Indigo works and demonstrate how it can facilitate interoperability between SVG and MusicXML (perhaps with the world premier of an original animated score in honor of this year's conference theme).

Our presentation requires only the most rudimentary knowledge of musical notation and SVG.

Bibliography

MakeMusic (2016). *MusicXML: Tutorial.* http://www.musicxml.com/tutorial/the-midi-compatible-part/ (accessed 27 Oct 2016).

Git-Lit: an Application of Distributed Version Control Technology toward the Creation of 50,000 Digital Scholarly Editions

Jonathan Reeve
jonathan.reeve@columbia.edu
Columbia University, United States of America

Distributed version control technologies, the most popular protocols of which are git, subversion, and mercurial, have long been popular among computer programmers for their abilities to track changes in a codebase and foster collaboration among coders. When combined with code management platforms such as GitHub, Bitbucket, or GitLab, they become even more powerful, enabling sophisticated bug tracking, project planning, and open-source code publication. Although these technologies have not yet been in widespread use in the humanities, their potential for use with corpus creation and textual editing is far-reaching. This paper describes Git-Lit, an open-source, community-centered initiative to parse, version control, and publish to GitHub roughly 50,000 scanned public-domain books from the British Library, thereby facilitating decentralized, open-access, and democratic scholarly editing.

The Git-Lit initiative addresses these problems:

- **Electronic texts are difficult to edit.** Traditional text repositories like Project Gutenberg and the Oxford Text Archive maintain central, canonical versions of their texts that, in most cases, are virtually immutable. If a reader spots an OCR error in an ebook, he or she must rely on contacting the publisher to propose a correction. Even with an infrastructure such as Project Gutenberg's Distributed Proofreaders, the process of releasing a corrected edition may take months or years. Git-Lit aims to radically streamline the improvement of an electronic text in two ways. First, ease of editing is achieved through GitHub's push-button forking (making a copy of a repository in one's user account) and in-browser editing---a reader may spot a mistake, correct it, and submit a pull request for the change in mere seconds, all without leaving the browser. Second, the decentralized model ensures that no single text may be considered unquestionably canonical, although *de facto* canonicity might be democratically achieved through repository voting mechanisms such as GitHub's stars.
- **Electronic texts often lack editorial history.** Owing, in some cases, to the age of an electronic text, its editorial provenance is often lost. Many Project Gutenberg editions, for instance, are transcribed from unknown print editions, and the history of their revisions is similarly opaque. Version control mitigates these problems by recording every edit, editor, and edition in the history of the text. When two editions diverge, git provides sophisticated tools for analyzing the differences between these editions. Sites like GitHub further provide graphical network charts, showing the genealogy of each version. Since contributions to a given text are logged according to individual contributor, credit for a given edition may be assigned according to the individual's percentage of total contributions, minimizing the danger, for instance, that a professor may take credit for his or her graduate student's work.
- **Textual corpora are difficult to assemble.** With some exceptions, notably the download function of the NLTK corpus module, downloading a text corpus involves compiling texts from diverse and heterogeneous sources. A would-be text analyst must click through a sequence of web pages to find the corpus he or she wants, and then either download a number of .zip files, or email the corpus creator to request a copy. With multiple texts, this can be a labor-intensive process that is not easily scriptable or automated. Git provides an easy way to solve these problems: by making texts available through the git protocol on GitHub, anyone that wishes to download a text corpus can simply run the command git clone followed by the repository URL. Parent repositories can then be assembled for collections of texts using git submodules. A parent corpus repository might be created for nineteenth-century *Bildungsromane*, for instance, and that repository would contain pointers to individual text repositories. These categories would not necessarily be mutually exclusive, and would allow for arbitrary curation of custom corpora. This provides a major advantage over the traditional directory structure model, where the existence of overlapping datasets necessitates the storage and maintenance of redundant data.
- **ALTO XML is not comfortably human-readable.** ALTO XML, the OCR output format used by the British Library texts, as well as texts created by the Library of Congress, is extremely verbose. It encodes the location of each word on the page, and often gives the OCR certainty for each word. While this format is useful for archival purposes, plain text editions are more useful for reading and for most brands of computational text analysis. Git-Lit parses the British Library's ALTO XML, and creates markdown versions of each text that are easily edited. Since markdown is readily converted to other document formats using tools such as Pandoc, this allows each text to be easily exported to PDF, EPUB, LaTeX, Docbook, and others. Additionally, Git-Lit is currently working on a system that leverages GitHub's built-in Jekyll HTML compilers to convert each text into a web page hosted on github.io, effectively creating 50,000 readable web editions. These new editions will exist as git branches in parallel with the markdown and ALTO XML editions. Since git maintains efficient copies of every historical version of the text, no information about the text is lost in these conversions. Anyone that wishes to improve the conversion script and create newer, better editions of the

original files may freely do so by branching the text from an earlier git commit.

Git-Lit software works by first parsing the XML metadata included with each text. This metadata is used to programmatically generate a repository name and a README.md file that describes the text, a document which GitHub will automatically render into a web page at the repository root. This file, along with standard CONTRIBUTING and LICENSE files, is then committed to local git repositories, initiating version control of the texts. The resulting local repository is then uploaded to GitHub via Python bindings to the GitHub API. Parent repositories are then created using git submodules for each collection of texts based on the their associated Library of Congress subjects. This enables a text analyst interested in 19th century poetry, for instance, to download all of the British Library's released works in this genre simply by running git clone https://github.com/git-lit/19th-century-poetry.git && git submodule update --init --recursive.

Since British Library texts are not the only ones being published to git-based platforms like GitHub---notable version-controlled corpora on GitHub include texts from the Text Creation Partnership and the early modern corpus Shakespeare His Contemporaries---git provides a common protocol for sharing, modifying, and distributing texts and textual corpora. Anyone may aggregate these corpora into parent repositories using git submodules. The Git-Lit project will soon launch a web application that will routinely scrape GitHub and other open repository sites for any textual corpus, thereby automating the process of discovering and indexing available corpora. This mechanism will also serve to democratize the curation of corpora, since the corpus index will be sorted by the number of GitHub "stars", or votes, a repository has engendered from the community.

The 50,000 British Library texts processed by Git-Lit, as well as many of the other open corpora described here, are currently being integrated into DHBox, the cloud-based Digital Humanities software suite. Soon, these corpora and many others will be available for download by selecting them from a web-based interface, where they will then be available for analysis using pre-installed versions of the Python NLTK, R, and other textual analytic tools.

This paper discusses how Git-Lit's methods might be used by other digital humanities projects involved in the creation or analysis of large text corpora, and how digital humanists may contribute to the Git-Lit project. (As an open-source project, Git-Lit welcomes contributions in the form of bug reports, feature requests, or code.) The paper also discusses some of the storage and computation limitations of electronically publishing texts via code repositories, and some of the technical problems encountered by the Git-Lit project. Finally, it suggests pedagogical uses of git-based collaborative digital editing, such as classroom compilation of anthologies or digital scholarly editions. The applications of these technologies are wide-ranging, and are neither proprietary to this project nor to services such as GitHub, but remain concepts of openness and collaboration with powerful implications for the digital humanities.

Researchers' perceptions of DH trends and topics in the English and Spanish-speaking community. DayofDH data as a case study.

Antonio Robles-Gómez
arobles@scc.uned.es
Spanish University for Distance Education, UNED

Elena González-Blanco
egonzalezblanco@flog.uned.es
Spanish University for Distance Education, UNED

Salvador Ros
sros@scc.uned.es
Spanish University for Distance Education, UNED

Gimena Del Rio Riande
gdelrio.riande@gmail.com
CONICET, Universidad de Buenos Aires

Roberto Hernández
roberto@scc.uned.es
Spanish University for Distance Education, UNED

Llanos Tobarra
llanos@scc.uned.es
Spanish University for Distance Education, UNED

Agustín C. Caminero
accaminero@scc.uned.es
Spanish University for Distance Education, UNED

Rafael Pastor
rpastor@scc.uned.es
Spanish University for Distance Education, UNED

Defining the "state of the art" in Digital Humanities (DH) is a really challenging task, given the range of contents that this tag covers. One of the most successful efforts in this sense has been the international blogging event known as "DayofDH" or "A Day in the Life of the Digital Humanities" project, promoted and sponsored by centerNet (http://www.dhcenternet.org/), which has put together digital humanists from around the world to document once a year what they do (Rockwell et al., 2012). The websites of DayofDH were hosted in North America until 2015, when it was coordinated in Europe by LINHD (http://linhd.uned.es), the Digital Innovation Lab, at UNED in Madrid. Participants belong to several countries around the world.

The relevance of DH in non-English speaking countries has been quick and important in the last decade, and especially important in the Spanish-speaking world (Spence and González-Blanco, 2014; González-Blanco, 2013; Del Rio Riande, 2014a; Del Rio Riande, 2014b; Galina et al., 2015). Technological projects for humanities have existed in the Spanish world for many years; however, the discipline called "Digital Humanities" arose in 2011 with the first meeting that originated the Spanish Digital Humanities Association, HDH. This relevance is reflected in the creation of a parallel version of the DayofDH in Spanish, the "DíaHD", which was hosted by the UNAM in Mexico in 2013 and 2014 and converged in the last initiative at UNED transforming both blogging events into a bilingual version of the Day.

Although there have been general studies about the information on participation in those events (Priani et al., 2014), there has not been an automated data analysis using NLP (Natural Language Processing) or Big Data tools to extract and classify the relevant information gathered in blogs (Webb et al., 2004). More technical details about these aspects can be found in (Tobarra et al., 2014b).

According to this, the main goal of this paper is to develop a dashboard that allows us to get more insight about interest topics and leaderships of this community during the period of time in which this event has been developed. With the "dashboard" word, we mean the analysis and presentation of results, not a tool. In this sense, the topic characterization process deals with the detection of the most relevant topics which are employed in the publication tools of these kinds of virtual communities (Tobarra et al., 2014a).

In order to achieve our aforementioned objectives, this work is focused on the datasets corresponding to four years of DayofDH (2012, 2013, 2014 and 2015 editions), and the Spanish version of the event in DíaHD 2013. This work strives at showing the evolution and trends in the last four years in order to give account of the presence of the Hispanic communities in the field. The information of the Spanish 2014 edition has been discarded, as it is not any more available online due to technical problems at the organizing institution. All editions of DayofDH employ WordPress, which has an associated SQL database, including several general tables and a specific set of tables per blog, defined in the project and common to all editions. The CMS is combined with the Buddypress social plugin, which lets users register, create communities and forums and interact among them. For the last edition of the Day, LINHD included also the bilingual plugin WPML to make it available the possibility of including translation in Spanish and English. This feature was, however, just used for the general website and its blog entries.

The data employed in this proposal has been obtained by using web scraping techniques (Fredheim, 2014) in the DayofDH websites for the previously mentioned editions. In particular, humanists' blog data, and their associated posts in the website have been gathered in this phase. All information scrapped from blogs is public and accessible from the Internet and, also, they have been anonymized for ethical issues. For validation purposes, the conceptual information about the database schemas have been compared with the extracted dataset, concluding the extraction process has been satisfactory.

Since the data obtained are huge enough to be efficiently processed, the use of big data techniques have been considered for this work (http://social-metrics.org/analyzing-big-data-with-python-pandas/). In order to achieve the main goals of the project, all the information related to the textual content in DayofDH have been processed, so that the most significant tokens are selected. Then, these tokens have been characterized by two parameters; first, it has been used the direct frequency which characterize if a token is used regularly in all DayofDH blogs. Secondly, the inverse frequency of the token that give information of how significant the token is in the context of digital humanities in a semantic way.

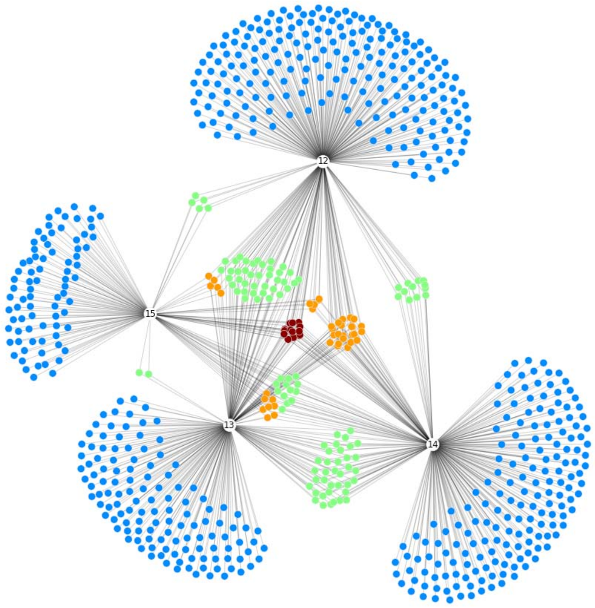


Fig. 1. Social network generated for 2012-2015 editions of DayofDH

These parameters have been used to observe the interest and evolution of the characterized tokens along the time, either in a global and individual way. The interest of the global analysis is to find how the knowledge has evolved during the years of the study. From the point of view of a personal analysis, the interest is to build individual profiles that show the main interest of the researchers in the humanist community. Finally, the leadership relations have also been explored by using disease propagation techniques in the generated social network, taken into account the different editions of the DayofDH. For instance, Fig. 1 shows the social network according to the amount of authors' participations.

The resulting graphics and visualizations (Tobarra et al., 2014b) let users make a quick idea of how the DH focus has been moving and distributing across the time through the different Academies in the different countries, but also how topics and interests change from one country to another and it is strongly related to their perspectives and disciplines, which are not independent from their origin (as an example, see Figs. 2 and 3). This approach will enlighten future studies on DH perspectives with real and precise data on the current state-of-the-art on DH perception and its evolution. Data of the years 2009, 2010 and 2011 are not used at the moment, as the same information is not available through web scraping. They will be incorporated to this study as a future work.

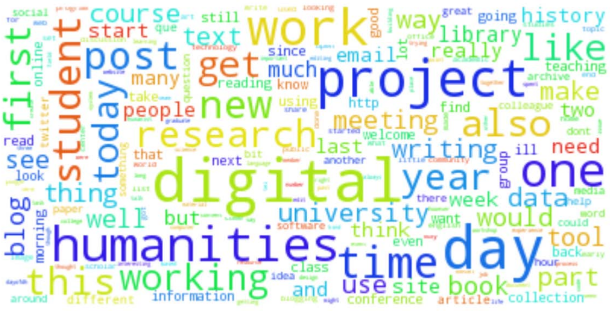


Fig. 2. Interest topics for 2012-2015 editions of DayofDH

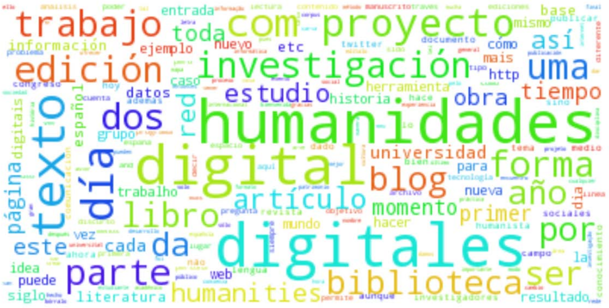


Fig. 3. Interest topics for 2013 edition of DíaHD

Acknowledgements

This work is part of next research projects funded by MINECO and leaded by the professor Elena González-Blanco; Research Europe Action EUIN2013-50630: Digital index of European poetry (DIREPO) and FFI2014-57961-R; and the Digital Humanities Innovation Lab: Digital edition, Linked data, and Research virtual environment for Humanities, in addition to the European project ERC-2015-STG-679528 POSTDATA. Authors would also like to acknowledge the support of the research project (2014I/PPRO/031) from UNED and Banco Santander, and a local project (2014-027-UNED-PROY). Furthermore, we thank both the Region of Madrid for the support of E-Madrid Network of Excellence (S2013-ICE2715) and the Spanish Government for the support of SNOLA Network of Excellence (TIN2015-71669-REDT).

Bibliography

Fredheim, R. (2014). Web-scraping: The basics. Available online at http://www.r-bloggers.com/web-scraping-the-basics/ (Accessed 9th February, 2016).

Galina, I., González-Blanco, E. and Rio Riande, G. del (2015). Se habla español. Formando comunidades digitales en el mundo de habla hispana (in Spanish). *Abstracts of the HDH 2015 Conference*. Madrid, Spain. Available online at http://hdh2015.linhd.es/ebook/hdh15-galina.xhtml (Accessed 9th February, 2016).

González-Blanco, E. (2013). Actualidad de las humanidades digitales y un ejemplo de ensamblaje poético en la red: ReMetCa (in Spanish). *Cuadernos Hispanoamericanos*, **761**: 53–67.

Priani, E., Spence, P., Galina, I., González-Blanco, E., Alves, D., Barrón Tovar, J. F., Godínez Bustos, M. A. and Paixão De Sousa, M. C. (2015). Las humanidades digitales en español y portugués. Un estudio de caso: DíaHD/DiaHD (in Spanish). *Anuario Americanista Europeo*, **12**: 5–18.

Rio Riande, G. del (2014a). ¿De qué hablamos cuando hablamos de humanidades digitales?. *Abstracts of the AAHD Conference. Culturas, Tecnologías, Saberes*. Buenos Aires, Argentina.

Rio Riande, G. del (2014b). ¿De qué hablamos cuando hablamos de Humanidades Digitales?. Available online at http://blogs.unlp.edu.ar/didacticaytic/2015/05/04/de-que-hablamos-cuando-hablamos-de-humanidades-digitales/ (Accessed 9th February, 2016).

Rockwell, G., Organisciak, P., Meredith-Lobay, M., Ranaweera, K., Ruecker, S. and Nyhan, J. (2012). The design of an international social media event: A day in the life of the digital humanities. *Digital Humanities Quarterly*, **6**(2).

Spence, P. and González-Blanco, E. (2014). A Historical Perspective on the Digital Humanities in Spain. *The Status Quo of Digital Humanities in Europe, H-Soz-Kult*.

Tobarra, Ll., Robles-Gómez, A., Ros, S., Hernández, R. and Caminero, A. C. (2014). Analyzing the students' behavior and relevant topics in virtual learning communities. *Computers in Human Behavior (CHB)*, **31**: 659–69.

Tobarra, Ll., Ros, S., Hernández, R., Robles-Gómez, A., Caminero, A. C. and Pastor, R. (2014). An integrated analytic dashboard for virtual evaluation laboratories and collaborative forums. *Tecnologias Aplicadas a la Ensenanza de la Electronica (Technologies Applied to Electronics Teaching) (TAEE), 2014 XI*. Bilbao, Spain, pp. 1–6.

Webb, E., Jones, A., Barker, P. and van Schaik, P. (2004). Using e-learning dialogues in higher education. *Innovations in Education and Teaching International*, **41**(1): 93–103.

Mnemosyne: A Smartlibrary for Rare and Forgotten Texts

Dolores Romero-López
dromero@filol.ucm.es
Universidad Complutense de Madrid, Spain

José Luis Bueren-Gómez-Acebo
jlbueren@hotmail.com
Universidad Complutense de Madrid, Spain

In recent decades European libraries have taken a giant step towards the mass digitization of their historical collections and the opening of their contents for the use of the global digital society. However, researchers and teachers experience great difficulties using, enriching or sharing that content. Our project aims to explore the new needs of the users of European digital libraries, databases and repositories in order to evolve the "traditional digital model" towards the *SmartLibrary* model, which proposes the compilation, integration and downloading of contents according to the needs of users and in order to enrich the European uses of the history.

Mnemosyne, for the ancient Greeks, was the personification of Memory. In our project the concept of memory comprises two uses:

1. **The recovery of historical memory** through European texts we consider rare and forgotten. We will analyse these concepts as conceptual categories in cultural studies (Alonso, 2008; Romero Lopez, 2014). This new paradigm will cover the analysis of a large and complex network of literary manifestations. We aim to record the history of the losers, looking for it in popular and mass culture texts that have been marginalized until now (Labanyi, 2003). This recovery requires making those digitized texts accessible and bringing together their interpretations so that the axes that have governed their oblivion within European cultures can be underlined.

2. **The rewriting of historical memory**. Once we have compared texts digitized in the *Scriptorium*, researchers and teachers will begin an enrichment of these texts through their collaborative annotation. This reinterpretation will allow historical memory to be restored by setting new categories of knowledge for the understanding of European cultures under common tendencies.

The prefix **SMART**- has been used as synonymous to agility, safety, ecology and sharing (Doran, 1981). It is a prefix that has been applied to phones, cars, houses and cities. So far, it has not been applied to libraries. The creation of a European *SmartLibrary* implies:

- **S**imple access to integrated European databases on contemporary and alternative European literatures
- **M**otivational search based on specific research content or didactic objects
- **A**ppropriate results based on the semantic Web search
- **R**esults discharged in a personal *Scriptorium* to be enriched with the specific tools
- **T**ransference of new digital objects to be shared with the global community

As a "smart digital model", exportable to other areas of the digital humanities, we have been developing *Mnemosyne: A SmartLibrary for Rare and Forgotten Texts*, based on the research that the LEETHI, LOEP and ILSA (see below) research groups in the Faculties of Philology and Information Technology of the Complutense University of Madrid (Spain) are jointly developing. Thanks to the collaboration of specialists in different European literatures and computer experts in the course of several national research projects, we have designed a new model of *Scriptorium* which allows for the integration of metadata and the enrichment of digital objects with new tools such as *Clavy* –an import/export tool for metadata- and *@Note* – a collaborative annotation tool.

About *Mnemosyne. Digital Library of the Other Silver Age* (Beta version):

1. *Mnemosyne. Digital Library of the Other Age of Silver* **is already accessible** on the Internet (http://repositorios.fdi.ucm.es/mnemosine/). As you can see *Mnemosyne* contains authors' data, access to digitized works and research collections. The field of study is rare and forgotten Spanish literary texts (1868-1939). The work is still in progress. Our current project ends by 06/31/2016.

2. *Mnemosyne records* show the metadata imported *by our tool Clavyfrom Biblioteca Digital Hispánica and from HathiTrust with the support of theComplutense Library*.

3. The *Mnemosyne* database works as **a laboratory** in which we experiment with *Clavy* the importation/exportation of metadata, and the tool *@Note* and practise collaborative annotation. *@Note* promotes the collaborative creation of free-text and semantic annotation schemas on literary works by communities of researchers, teachers and students and the use of these schemas in a very flexible and adaptive model for the definition of annotation activities.

4. As *SmartLibrary, Mnemosyne* **will integrate** @Note **and other digital** tools (forthcoming). Of course we are very much interested in DARIAH tools and its research infrastructure and we would like to collaborate with this European consortium. Besides, our *SmartLibrary* could be an extraordinary field of study in which to experience the development and integration of new digital tools. *Mnemosyne*, as a *SmartLibrary*, could become a field of international experimentation for the practice of tools with semantic interoperable networks.

5. The Spanish *Mnemosyne* is the first example of what we would like to build. We would like to regrow our *smart* model in the international environment with the support of other European projects, interested, like the authors, in rare and forgotten texts and the uses of the past.

Mnemosyne: SmartLibrary for Rare and Forgotten Texts **needs transnational collaboration**. This project involves the integration of specialists in different European literatures and researchers in computer science to develop new research and resources for the common use of European citizens.

The research developed in *Mnemosyne* is being financed by: 1) Ministerio de Economía y Competitividad. Research Project: "Escritorios Electrónicos para las Literaturas-2". Reference FFI2012-34666 (2012-2016). Directora: Dolores Romero López, Facultad de Filología, Complutese University of Madrid . 2) I Convocatoria de Ayudas a Proyectos de Investigación de la Fundación BBVA: "Modelo unificado de Gestión de Colecciones Digitales con Estructuras Reconfigurables: Aplicación a la Creación de Bibliotecas Digitales Especializadas para Investigación y Docencia". Reference: HUM14_251(2015-2016). Director: José Luis Sierra Rodríguez, Facultad de Informática, Complutense University of Madrid

RESEARCH GROUPS: LEETHI Research Group, ILSA Research Group, LOEP Research Group.

Bibliography

Alonso, C. (2008). Sobre la categoría canónica de 'raros y olvidados'. *Anales de Literatura Española*, **20**: 11-38.

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review* (AMA FORUM), **70**(11): 35–36.

Labanyi, J. (2003). *Constructing Identity in Contemporary Spain*. Oxford: Oxford University Press.

Romero López, D. ed. (2014). *Los márgenes de la modernidad. Temas y creadores raros y olvidados en la Edad de Plata*. Sevilla: Punto Rojo Libros.

Romero López, D. (2014). Hacia la SmartLibrary: Mnemosyne, una biblioteca digital de textos literarios raros y olvidados de la Edad de Plata (1868-1936). Fase I, en *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, Sagrario López Poza y Nieves Pena Sueiro (editoras), *Janus*, Anexo **1**: 411-22.

Romero López, D. (2015). *Bibliotecas digitales inteligentes para la docencia y la investigación*. Eprints.ucm.es/31422/1/MySmartLibrary_Escorial_DRomero_2015.pdf.

Some Problems in the Non-Traditional Authorship Attribution Studies of the Dramatic Canon of William Shakespeare: Are they Insurmountable

Joseph Rudman
jr20@heps.phys.cmu.edu
Department of English, Carnegie Mellon University, United States of America

William Shakespeare is arguably the greatest dramatist of all time. Yet, the man and his works are shrouded in mystery and uncertainty. This paper posits that a man named Shakespeare wrote the First Folio. It is this First Folio that provides the "most certain" body of Shakespeare's plays.

The makeup of the Shakespeare dramatic canon has prompted more attribution studies (traditional and non-traditional) and caused more controversies than any other canon – by far. This paper looks at the several hundred non-traditional studies (and concomitant "flame wars") and points out some of the more serious problems. There is no doubt that it is a canon in disarray. Most of the scholars involved in the controversies are the "heavyweights" of Shakespearean studies – e.g.:

- Rasmussen (1977) vs. Hope (1994)
- Vickers (2011) vs. Craig and Kinney (2009)
- Taylor (2015) vs. Stern (2004)
- Craig vs. Vickers and Jackson (Hirsch and Craig, 2014)

But the most famous controversies involve (1) Donald Foster vs. Ward Elliott and (2) Robert Valenza and Donald Foster vs. the world.

In a recent article (Rudman, 2016) I pointed out many caveats to scholars working on authorship attribution on the canon of William Shakespeare. Among these are:

- Reproducibility
- Input Texts
- Genre
- Editing
- Controls
- Isolation of Variables
- Choice of Style Markers
- Statistical Tests
- Sample Selection and Size
- Treatment of Errors
- Collaboration

In this paper, I expand on one of these caveats (Genre), look at and cite examples from many of the non-traditional studies of the Shakespeare canon in order to highlight these problems, and suggest solutions.

- Each genre is governed by different linguistic rules and rhetorical purpose – a practitioner should not mix genre.

- Drama – Comedy, History, Tragedy, Romance – how far down should these be catagorized
- Verse within the drama – rhymed verse within the verse – how far down should this be broken
- Music
- Dialogue vs Dramatic Monologue

Also in this paper, I address the conundrum of using what I deem as seriously flawed studies to show problems with other studies – e.g. if a practitioner mixes genres (history and tragedy) but shows that Shakespeare's style changes over time, I cite this change as evidence that chronological constraints must be employed.

No matter how flawed I consider a study, there are parts of that study that are correct – there is no non-traditional study (of the hundreds conducted) that is completely without merit. By invoking the *etiam si non est verum* paradigm, I show how almost all of the necessary steps in a valid study exist in the literature. We can look at all of the links in he chain (even the broken ones) and try to piece together what we should do to move the field forward. The evidence in this paper reinforces my conclusion from the *JEMS* article that any attribution results are problematic at best.

The Conclusion reached (and I believe supported by very strong evidence) is that all of the non-traditional studies are seriously flawed and that as of today we do not have a valid Shakespeare text to conduct non-traditional attribution studies.

The following bibliography is only representative. There are more than 200 non-traditional studies on the Shakespeare canon. My working bibliography for this study is well over 1,000 entries.

Bibliography

Bullough, G. (1957). *Narrative and Dramatic Sources of Shakespeare*, Vol. **2**(8), London: Routledge and Kegan Paul.

Burrows, J. (2012). A Second Opinion on Shakespeare and Authorship Studies in the Twenty-First Century. *Shakespeare Quarterly*, **63**(3): 355–92.

Busse, U. (2002). *Linguistic Variation in the Shakespeare Corpus: Morpho-syntactic Variability of Second Person Pronouns*. Amsterdam–Philadelphia: John Benjamins.

Hirsch, B. D. and Craig, H. (2014). Mingled Yarn: The State of Computing in Shakespeare 2.0. In Bishop, T. and Huang, A., Hirsch, B. D. and Craig, H. (eds.), *The Shakespearean International Yearbook*. Aldershot: Ashgate Publishing, preprint edition courtesy of publisher, **14**: 3-35.

Hope, J., and Whitmore, M. (2014). Quantifiaction and the Language of Later Shakespeare. *Actes des Congrès de la Société Française Shakespeare*, **31**: 123–49.

Horton, T. B. (1987). *The Effectiveness of the Stylometry of Function Words in Discriminating Between Shakespeare and Fletcher*. Ph.D. Thesis, University of Edinburgh, 1987.

Lancashire, I. (2002). The State of Computing in Shakespeare. In Elton, W. R. and Mucciolo, J. M. (eds.), *The Shakespearean International Yearbook. Where are We Now in Shakespearean Studies?*. Aldershot: Ashgate Publishing, **2**: 89–110.

Mowat, B. A. (2003). Whats in a Name? Tragicomedy, Romance, or Late Comedy. In Dutton, R. and Howard, J. E. (eds.), *A Companion to Shakespeare's Works. The Poems, Problem Comedies, Late Plays*. Oxford: Blackwell Publishing, **4**: 129–49.

Rudman, J. (2016). Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats. *Journal of Early Modern Studies*, forthcoming.

Vickers, B. (2009). *Shakespeare and Authorship Studies in the Twenty-First Century. Review Essay. Shakespeare Quarterly*, **62**(1): 106–42.

Witmore, M. (2009). *A Genre Map of Shakespeare's Plays from the First Folio. Wine Dark Sea*, http.//www.winedarksea.org/?p=40, (Accessed May 6, 2015).

Climate Negotiation Analysis

Pablo Ruiz Fabo
pabloruizfabo@gmail.com
LATTICE Lab, École Normale Supérieure, France

Clément Plancq
clement.plancq@ens.fr
LATTICE Lab, École Normale Supérieure, France

Thierry Poibeau
thierry.poibeau@ens.fr
LATTICE Lab, École Normale Supérieure, France

Introduction

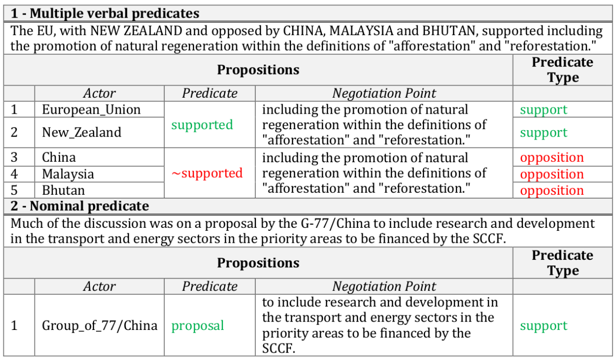
Text analysis methods based on word co-occurrence have yielded useful results in humanities and social sciences research. For instance, Venturini et al., (2012) describe the use of concept co-occurrence networks in social sciences. Grimmer and Stewart (2013) survey clustering and topic modeling applied to political science corpora. Whereas these methods provide a useful overview of a corpus, they cannot determine the predicates[1] relating co-occurring elements with each other. For instance, if *France* and the phrase *binding commitments* co-occur within a sentence, how are both elements related? Is France in favour of, or against *binding commitments*?

Different natural language processing (NLP) technologies can identify related elements in text, and the predicates relating them. A recent approach is *open relation extraction* (Mausam et al., 2012, among others), where relations are derived from the corpus in a data-driven manner, without having to pre-specify a vocabulary of predicates or actors. We are developing a workflow to analyze the Earth Negotiations Bulletin (vol. 12)[2], which summarizes international climate negotiations. A sentence in this corpus can contain several verbal or nominal predicates indicating support and opposition (see Table 1). Results were uneven when applying open relation extraction tools to this corpus. To address these challenges, we developed a workflow with a domain model, and analysis rules that exploit annotations for semantic roles and pronominal anaphora, provided by an NLP pipeline.

Our system identifies points supported and opposed by negotiating actors and extracts keyphrases and DBpedia[3] concepts from those points. The results are displayed on an interface, allowing for a comparison of different actors' positions. The system helps address a current need in digital humanities: tools for the quantitative analysis of textual structures beyond word co-occurrence.

The abstract is structured as follows. First, related work and the corpus are presented. Then, our system is described. Finally, evaluation is discussed.

Material supplementing the paper and access information to the system will be available on the project's website.[4]



1 - Multiple verbal predicates				
The EU, with NEW ZEALAND and opposed by CHINA, MALAYSIA and BHUTAN, supported including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."				
Propositions				**Predicate Type**
	Actor	*Predicate*	*Negotiation Point*	
1	European_Union	supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	support
2	New_Zealand			support
3	China	~supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	opposition
4	Malaysia			opposition
5	Bhutan			opposition
2 - Nominal predicate				
Much of the discussion was on a proposal by the G-77/China to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.				
Propositions				**Predicate Type**
	Actor	*Predicate*	*Negotiation Point*	
1	Group_of_77/China	proposal	to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.	support

Table 1: **Typical corpus sentences**. Sentence 1 has predicates *supported* and *opposed*, with several actors each. Example 2 shows a nominal predicate (*proposal*). For Sentence 1, five ‹*actor, predicate, negotiation point*› propositions are extracted by the system, and the opposing actors (*China, Malaysia, Bhutan*) are assigned a proposition which is a negated version (with *~supported* as the predicate) of the proposition for the main verb *supported*.

Related work

Venturini et al., (2014) created concept co-occurrence networks for the ENB corpus, using Cortext Manager[5], a corpus cartography tool. This analysis does not cover which predicates relate concepts and actors. Salway et al., (2014) used *grammar induction* on ENB to identify recurrent actor/predicate patterns; it could be tested whether results with that approach complement ours.

Some studies have used syntactic and semantic parsing for text-analysis of social sciences and humanities corpora. Diesner (2012, 2014) examines the contribution of NLP to the construction of text-based networks. Van Atteveldt (2015) used dependency parsing to apply co-occurrence based methods within sentence elements related to an actor or a predicate. These studies rely mostly on syntactic dependencies and verbal predicates. We are using semantic role labeling as the basis for relation extraction, and treating nominal predicates besides verbal ones. We also developed an interface to navigate the results.

Finally, a relevant resource for text-mining on climate corpora is *climatetagger* API[6], which links concepts against a domain-specific thesaurus (Bauer et al., 2011). This thesaurus could complement our concept-linking results (based on DBpedia, a general ontology).

Corpus

Each ENB issue is a 2000 word summary for one day of negotiations. The issues are written by domain experts, who strive for an objective tone and, to avoid biases, use similar expressions when reporting about all participants' interventions (Venturini et al., 2014). The COP meetings are covered in 255 ENB issues, with ca. 35,000 sentences. The original corpus format is HTML, which we preprocessed into clean text. We dated each issue based on ENB's table of contents.

System description

The system helps analyze patterns of support and opposition between negotiating parties, and the issues about which parties agree or disagree. To achieve this, the system extracts propositions of shape ‹*actor, predicate, negotiation point*›,[7] based on a domain model containing actors and predicates, and applying analysis rules on the outputs of an NLP toolkit. Keyphrases and DBpedia concepts are also extracted from the negotiation points. All extractions, and the corpus itself, are made navigable on a user interface (UI).

NLP toolkit

We used the IXA Pipes library[8] (Agerri et al., 2014), with default models for **tokenization** and **part-of-speech tagging**. We resolved some types of **pronominal anaphora** based on *CorefGraph*[9] coreference chains.

Semantic Role Labeling (SRL) (Surdeanu et al., 2008) identifies a predicate's arguments and their semantic functions or roles (e.g. *agent*). SRL was performed with ixa-pipe-srl[10], which tags against the PropBank database (Palmer et al., 2005) for verbal predicates and against NomBank (Meyers et al., 2004) for nominal ones.

Keyphrase Extraction: YaTeA[11] was used (Aubin and Hamon, 2006). This library performs unsupervised term extraction using syntactic and statistical criteria.

Entity Linking (EL): The tool from (Ruiz and Poibeau, 2015) was used. It combines outputs from several public EL services, selecting the best outputs with a weighted vote.

Domain-specific components

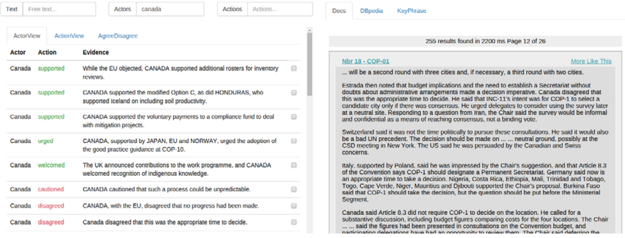
The **domain model** contains actors (negotiating countries and groups) and verbal or nominal predicates. Verbal predicates (from PropBank) can be neutral reporting verbs (e.g. *stated*), or verbs related to support and opposition (*recommended*, *criticized*). The nominal predicates (from NomBank) express similar notions to the verbs (e.g. *proposal*, *objection*). The model also specifies a predicate type: *report*, *support*, or *oppose*.

Analysis rules were implemented to identify propositions based on the semantic roles of predicates' arguments, previously obtained with SRL. Most domain predicates involve an agent and a message expressed by the agent (who agrees with the message, objects to it, or just reports it). Thus, actor mentions in a predicate's A0 argument[12] represent the actor who expresses the message, and the predicate's A1 argument 12 often represents the negotiation point addressed by the actor. The generic rule to identify propositions is in Figure 1.

```
Rule: Generic proposition

for each predicate p :
  · resolve negation (see below)
  for each pronoun he, she, it in p's A0 argument :
    · apply anaphora resolution (see below)
  for each actor-mention am in p's A0 argument :
    · create a proposition ‹am, p, point›, where point is a concatenation of p's A1 arguments
```

Figure 1: **General rule to create a proposition**

Sentences with *opposed by* constructions require a different analysis (e.g. *China, opposed by the EU, recommended…*) In such sentences, a different rule creates, for the opposing actors, propositions where the predicate contradicts the main clause's predicate (see Table 1 for an example). Proposition-creation rules for more specific cases have also been implemented.



Text | Free text...

Actors | canada

Actions | Actions...

Actor	Action	Evidence	
Canada	supported	While the EU objected, CANADA supported additional rosters for inventory reviews.	□
Canada	supported	CANADA supported the modified Option C, as did HONDURAS, who supported Iceland on including soil productivity.	□
Canada	supported	CANADA supported the voluntary payments to a compliance fund to deal with mitigation projects.	□
Canada	urged	CANADA, supported by JAPAN, EU and NORWAY, urged the adoption of the good practice guidance at COP-10.	□
Canada	welcomed	The UK announced contributions to the work programme, and CANADA welcomed recognition of indigenous knowledge.	□
Canada	cautioned	CANADA cautioned that such a process could be unpredictable.	□
Canada	disagreed	CANADA, with the EU, disagreed that no progress had been made.	□
Canada	disagreed	Canada disagreed that this was the appropriate time to decide.	□

255 results found in 2200 ms Page 12 of 26

Nbr 18 - COP-01 More Like This

... will be a second round with three cities and, if necessary, a third round with two cities.

Estrada then noted that budget implications and the need to establish a Secretariat without doubts about administrative arrangements made a decision imperative. Canada disagreed that this was the appropriate time to decide. He said that INC-11's intent was for COP-1 to select a candidate city only if there was consensus. He urged delegates to consider using the survey later at a neutral site. Responding to a question from Iran, the Chair said the survey would be informal and confidential as a means of reaching consensus, not a binding vote.

Switzerland said it was not the time politically to pursue these consultations. He said it would also be a bad UN precedent. The decision should be made on ... ... neutral ground, possibly at the CSD meeting in New York. The US said he was persuaded by the Canadian and Swiss concerns.

Italy, supported by Poland, said he was impressed by the Chair's suggestion, and that Article 8.3 of the Convention says COP-1 should designate a Permanent Secretariat. Germany said now is an appropriate time to take a decision. Nigeria, Costa Rica, Ethiopia, Mali, Trinidad and Tobago, Togo, Cape Verde, Niger, Mauritius and Djibouti supported the Chair's proposal. Burkina Faso said that COP-1 should take the decision, but the question should be put before the Ministerial Segment.

Canada said Article 8.3 did not require COP-1 to decide on the location. He called for a substantive discussion, including budget figures comparing costs for the four locations. The Chair ... ... said the figures had been presented in consultations on the Convention budget, and participating delegations have had an opportunity to review them. The Chair said deferring the

Figure 2: **Main view of the interface.** The left panel gives access to the search workflows (Actors, Actions, Points). It also shows propositions for a query (e.g. the actor *Canada*), and gives access to the *AgreeDisagree* view. The right panel shows the documents in the Docs tab, as well as aggregated keyphrases and DBpedia concepts for the query or for selected propositions, in the other tabs.

The treatment of **negation** relies on finding *AM-NEG* roles (see footnote 12) attached to a predicate, or negative items (*not*, *lack*) in a window of two tokens preceding a predicate.

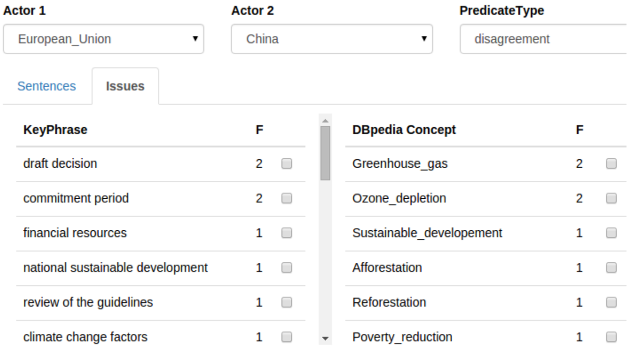
Pronominal anaphora was treated via custom rules operating on the output of a coreference resolver (see footnote 9). We created custom rules since, in the corpus, *he* and *she* (besides *it*) can refer back to a country (pronoun gender depends on the country's delegate).

To facilitate searches by date-range, propositions are assigned their documents' date.

User interface

The UI (Figure 2) helps analyze actors' negotiation positions. It allows searching for documents matching a text query (Text search box), and for propositions matching a given actor (Actors box) or a given predicate (Actions box). Propositions matching a query are displayed on the left panel, documents for a query on the right. Aggregated keyphrases and DBpedia concepts for the content matching a query (documents or propositions) are displayed in tabs on the right panel. The AgreeDisagree view provides an overview of keyphrases and concepts from propositions where selected actors agree or disagree. Simultaneous access on the UI to the corpus and the annotations helps researchers validate results.

The implementation framework is Django[13], with Solr search.[14] We're working on allowing the user export results and edit the model's actors and predicates.



Actor 1: European_Union ▾

Actor 2: China ▾

PredicateType: disagreement

KeyPhrase	F	
draft decision	2	□
commitment period	2	□
financial resources	1	□
national sustainable development	1	□
review of the guidelines	1	□
climate change factors	1	□

DBpedia Concept	F	
Greenhouse_gas	2	□
Ozone_depletion	2	□
Sustainable_developement	1	□
Afforestation	1	□
Reforestation	1	□
Poverty_reduction	1	□

Figure 3: **AgreeDisagree** View displays keyphrases and DBpedia concepts from propositions where actors (here the EU and China) agree or (as here) disagree.

Evaluation

It is important to assess whether the system can help domain-experts gain insights they would not have otherwise obtained, e.g. detect previously unnoticed generalizations (see e.g. Berry, 2012). This type of evaluation is ongoing; we are collaborating with political scientists, whose initial feedback on the tool has been positive. User validation of the interface is also ongoing.

The system's NLP components were evaluated in literature cited above. Results are state-of-the-art or competitive, and available on our project's website (sites.google.com/site/nlp4climate).

To evaluate the model and analysis rules that create domain-relevant propositions, we have manually annotated a set of corpus sentences with propositions. Details about the test-set, evaluation metrics and results are on the website. We consider the results satisfactory.

Outlook

A useful feature would be an annotation confidence score, that users could employ to establish priorities in manual result revision. A useful application of the propositions extracted would be creating network graphs with different types of edges representing support and opposition among parties, and between parties and issues.

Acknowledgements

We thank Tommaso Venturini, Audrey Baneyx, Kari de Pryck and Diégo Antolinos-Basso from the Sciences Po médialab in Paris for domain-expert feedback on the system. Pablo Ruiz is supported by a PhD grant from Région Ile-de-France.

Bibliography

Agerri, R., Bermudez, J. and Rigau, G. (2014). IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of LREC 2014, the 9th Language Resources and Evaluation Conference*. Reykjavik, Iceland.

Aubin, S. and Hamon, T. (2006). Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006*, LNAI 4139. Springer, pp. 380-87.

Auer, S., Bizer, C., Kobilarov, G., Lehman, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The Semantic Web*, Springer, pp. 722–35.

Bauer, F., Recheis, D. and Kaltenböck, M. (2011). Data. reegle. info–A new key portal for Open Energy Data. In *Environmental Software Systems. Frameworks of eEnvironment*, Springer Berlin Heidelberg, pp. 189-94.

Berry, D. M. (2012). *Understanding Digital Humanities*, pp. 1–20. Palgrave Macmillan.

Björkelund, A., Bohnet, B., Hafdell, L. and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Coling 2010, 23rd International Conference on Computational Linguistics: Demonstration Volume*, Beijing, pp. 33–36,

Diesner, J. (2012). *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts*. PhD Thesis. Carnegie Mellon University.

Diesner, J. (2014). ConText: Software for the Integrated Analysis of Text Data and Network Data. In *Social and Semantic Networks in Communication Research*, at *ICA*, *Conference of International Communication Association*.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, OUP, pp. 1–31.

Mausam, Schmitz, M., Bart, R., Soderland, S. and Etzioni, O. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–34.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, V. and Grishman. R. (2004). The NomBank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, pp. 24–31.

Palmer, M., Gildea, D. and Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, **31**: 1.

Ruiz, P. and Poibeau, T. (2015). Combining Open Source Annotators for Entity Linking through Weighted Voting. In *Proceedings of *SEM. Fourth Joint Conference on Lexical and Computational Semantics,* Denver, U.S., pp. 211–15.

Salway, A., Toulieb, S. and Tvinnereim, E. (2014). Inducing information structures for data-driven text-analysis. *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*, pp. 28–32.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, pp. 159–77.

Van Atteveldt, W., Sheaferm, T., Shenhav, S., and Fogel-Dror, Y. (2015). Clause analysis: using syntactic information to enrich frequency-based automatic content analysis. In *Symposium New Frontiers of Automated Content Analysis in the Social Sciences*, at the University of Zurich.

Venturini, T. and Guido, D. 2012. Once upon a text: an ANT tale in Text Analytics. *Sociologica,* **3**: 1–17. Il Mulino, Bologna.

Venturini T., Baya Laffite, N., Cointet, J-P., Gray, I., Zabban, V., and De Pryck, K. (2014). Three maps and three misunderstandings: A digital mapping of climate diplomacy. *Big Data and Society*,**1**(2): 1–19.

Notes

[1] *Predicate* in the sense of an expression relating a set of arguments.

[2] http://www.iisd.ca/vol12

[3] wiki.dbpedia.org (Auer et al., 2007)

[4] https://sites.google.com/site/nlp4climate

[5] http://docs.cortext.net

[6] API: http://api.climatetagger.net ; Thesaurus: http://www.climatetagger.net/glossary/

[7] Terminology adopted: ‹*Norway, preferred, legally-binding commitments*› is a proposition, with actor *Norway*, predicate *preferred* and *legally-binding commitments* as the negotiation point.

[8] http://ixa2.si.ehu.es/ixa-pipes/

[9] https://bitbucket.org/Josu/corefgraph

[10] https://github.com/newsreader/ixa-pipe-srl ; it provides a wrapper to *mate-tools* (Björkelund et al., 2010)

[11] http://search.cpan.org/~thhamon/Lingua-YaTeA/lib/Lingua/YaTeA.pm

[12] In SRL, *A0* corresponds to a predicate's agent. *A1* is the patient or theme. *AM* roles represent adjuncts (time, location etc.) or negation. See Palmer et al., 2005.

[13] https://www.djangoproject.com/

[14] https://lucene.apache.org/solr/

Modeling as Discourse: The case for 3D

Marie Giltner Saldaña
marie.saldana@gmail.com
University of North Carolina, Chapel Hill, United States of America

Introduction

In 2004, Willard McCarty characterized the Digital Humanities as “an experimental practice”, with modeling at its core: “the way to a computing that is *of* as well as *in* the humanities: a continual process of coming to know by manipulating representations” (McCarty, 2004). McCarty’s proposition embraced the open-endedness of models, yet this quality tends to be forgotten in the face of 3D’s sensory immediacy, with the result that such models tend to be perceived as static. Diane Favro, a longtime proponent of 3D modeling for architectural historical research, has acknowledged that “while observers intellectually acknowledge that the virtual re-creation is an approximation, not a *doppelgänger* for a past reality, this concept is almost immediately subsumed by the experiential power of the presentation” (Favro, 2006). This seems to hold true regardless of the model’s realism. Recent attempts to mitigate this effect emphasize the metadata and decision making processes behind 3D models, whose claim to scholarship is substantiated by virtue of their being supported by a database of textual, graphic, and quantitative sources (Saldaña, 2015). Anyone who has worked on 3D historical models, however, can attest that the act of compiling and layering of source data only reinforces the ultimate inability of such multiplicitous archives to present a coherent picture of the past on their own terms, let alone generate a truly representational or mimetic 3D model. Despite this, it is difficult to prevent even the database-driven 3D model from slipping back to the mode of uncritical simulation.

Simulation intentionally obscures the digital means of production in order to present a more convincing version of reality, in the service of a predetermined purpose. The origins of virtual reality simulation are coincident with the post-WWII military-industrial complex, with military training schemes being some of the primary instigators and funders of computer graphics simulation research (Penny, 2004). The anti-discursive nature of simulations is embedded in their interfaces, which invite the observer to experience the space in as natural a manner as possible, whether through embodied movement or the abstracted kinesis of mouse clicks. The pedagogical implications for academia are clear. VR training programs used in medicine, aviation, and the military rely on the repetitive, unthinking response of users to train them in a particular behavior (Penny, 2004). But what of the open-endedness of the humanistic modeling project? The automatic assumption that 3D space maps real space obscures the relationship of 3D modeling with critical theories of modeling practice in the humanities. Can 3D models be recouped from the totalizing logic of simulation?

This paper attempts to access this underexplored theoretical potential of 3D by proposing a reconceptualization of 3D modeling in the Foucauldian sense of *discourse,* rather than representation. Foucault’s idea of discourse was not meant to uncover an underlying definitive meaning of a given text or object, that elusive “knowledge” to which so many scholarly 3D models aspire. Rather, discourse is meant to uncover the “rules that are revealed when an object of discourse is modified and transformed”, which often involve implicit power structures (Philp, 2013). 3D digital models, which embed and incorporate many types of information, have the capacity to bring these layers of knowledge into dialogue with each other so that the underlying assumptions inscribed in them are revealed.

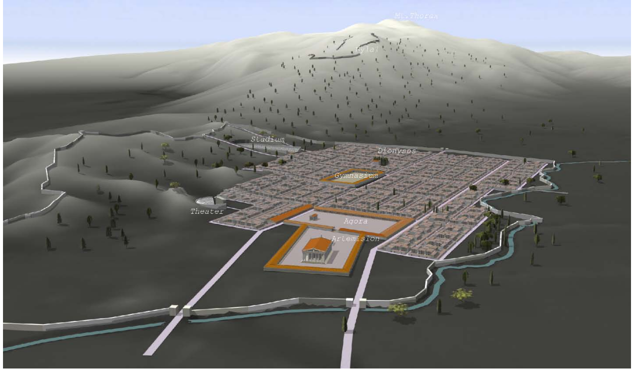


Fig.1 3D procedural model of the Hellenistic city of Magnesia on the Maeander

Case Studies

This paper will examine two of the author’s projects as ongoing attempts to mine the discursive power of 3D modeling in different ways. The first case study is an exploration of the vocabularies of ancient Roman architecture at Magnesia on the Maeander. The project directly engages with the concept of ‘rules’ that allow the formation of discourse as a “system of possibility for knowledge” (Philp, 2013). In the context of procedural modeling, the methodology used here, *rule* is a technical term denoting the scripts that generate 3D visualizations, and it also accurately describes the heuristic process of modeling. In writing the procedural rules, the modeler seeks to match a proleptic model with a mimetic one, continually negotiating the gaps between conjecture and representation. This process becomes discourse when the rules that result begin to delineate the shape of knowledge formation itself. Procedural modeling goes part of the way towards formulating a discursive approach to the process of modeling. As a finished product, however, 3D procedural models are indistinguishable from their non-discursive counterparts (Fig.1). Another possible answer involves dissolving the link between 3D space and representational space. The second case study is an application built in the Processing development environment that takes as its object of study the *Manhattan Transcripts* by Bernard Tschumi (Tschumi, 1981). Like the archaeological data behind the 3D city model of Magnesia, the *Manhattan Transcripts* comprises different representational modes to form a narrative that leaves many gaps open to interpretation. Originally displayed as an exhibition and later published as a book, Tschumi's work contains 24 sets of 3 images for the same event, expressed as plan, diagram, and rendering. The digital application attempts to bring these images into dialogue with each other through the medium of 3D space by separating them and recombining them in various ways (Fig.2). 3D space here serves the discursive function of exposing the images' areas of overlap and discordance in combinations and sequences not possible on the printed page or gallery wall. The sets of photographs, drawings, and movement diagrams can be viewed side-by-side or as semi-transparent layers thus making it easier to investigate the correspondences and differences between each representational mode. In addition, the images can be viewed all together or sorted by medium. Using the keyboard, the images can be cycled like frames in a film, highlighting Tschumi's own references to the medium of cinema.

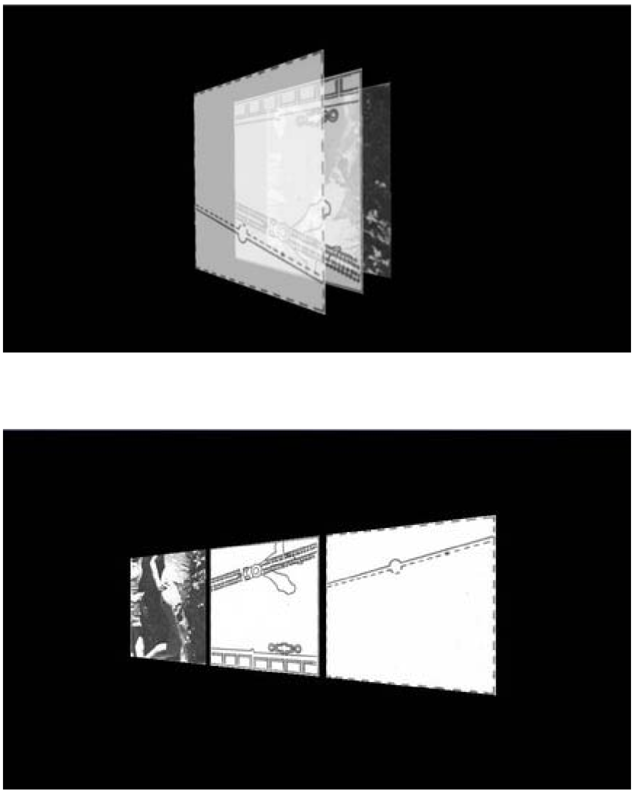


Fig. 2 Screenshots from the 'Manhattan Transcripts' 3D environment

Conclusions

The use of 3D media in the humanities has come of age in the last 15 years and its function as a critical apparatus is still in the process of development. Projects that treat 3D models as more than representational tools have the capacity to expand and illuminate this increasingly widespread form of digital culture. Particularly important is the need for theoretical discourse that resists the intentionality of 3D simulations and provides alternative modes of generative, non-determinant modeling. The case studies mentioned in this paper are but two partial answers to this challenge, part of a growing effort by a contingent of artists and researchers interested in the critical potential of 3D models.

Bibliography

McCarty, W. (2004). Modeling: A Study in Words and Meanings. In Schreibman, S., Siemens, R., and Unworth, J. (Eds.), *A Companion to Digital Humanities*. Oxford: Blackwell.

Penny, S. (2004). Representation, Enaction, and the Ethics of Simulation. *Electronic Book Review,* http://www.electronicbookreview.com/thread/firstperson/machanimate (accessed 5 March 2016).

Philp, M. (1990). Michel Foucault. In Skinner, Q. (Ed). *The Return of Grand Theory in the Human Sciences*. Cambridge: Cambridge University Press.

Saldaña, M. (2015). An Integrated Approach to the Procedural Modeling of Ancient Cities and Buildings. *Digital Scholarship in the Humanities,* 30, Suppl. **1**: 48-63.

Tschumi, B. (1981). *The Manhattan Transcripts*. New York: St. Martin's Press.

Mapping European Periodical Counterpublics: Building a Sustainable Collaborative Framework for European Periodical Studies

Jasper Schelstraete
jasper.schelstraete@ugent.be
Ghent University, Belgium

Sally Chambers
sally.chambers@ugent.be
Ghent University, Belgium

Marianne Van Remoortel
marianne.vanremoortel@ugent.be
Ghent University, Belgium

Agents of Change: Women Editors and Socio-Cultural Transformation in Europe, 1710–1920 is a five–year humanities research project funded by a European Research Council (ERC) Starting Grant (2015–2020), directed by Marianne Van Remoortel from the Department of Literary Studies at Ghent University, Belgium. It examines a neglected aspect of the social and cultural life in Europe in the modern period: the impact of women editors on public debate. From the 1700s on, European women actively participated in the cultural arena through the journals that they edited. *Agents of Change* advances the hypothesis that periodical editorship enabled women editors to take a prominent role in public life and as a result influence public opinion and shape transnational processes of socio-cultural change. By examining how these processes unfolded in the press through practices of textual transfer both among women and in the larger publishing landscape, *Agents of Change* will not only initiate a shift in our thinking about the participation of women in society and print culture, but also pave the way for pan-European research on the periodical press.

In order to trace these networks of intellectual exchange, *Agents of Change* is using NodeGoat, a web-based integrated data management, network analysis, and visualisation platform, developed by Lab1100, a research and development company, based in The Netherlands. NodeGoat allows us to collaboratively gather our research data about women editors and their periodicals, and enables us to visualise and analyse the linkages (both biographical and bibliographical) between them. By gathering evidence to prove connections between people and publications across languages and state borders, we will be able to identify the dynamics of cultural prestige at work in Europe. For example, how knowledge and fashion radiated outward from a few trendsetting periodicals across the pages of myriad publications which translated, adapted, or reprinted them either in part or in their entirety. The data that we collect in NodeGoat will also be invaluable as a descriptive index of periodical editors, a role which traditional print culture studies has tended to overlook, especially when it comes to women periodical editors. For this reason, we are developing a web front-end that will act as a catalogue interface to make this descriptive index freely available online.

In order to fully capture these transnational networks of intellectual exchange it is important to strive for the most comprehensive coverage possible of the period and region at hand. Our multilingual and multidisciplinary team of six researchers will pay particular attention to practices of textual transfer (including translation, adaptation, reprinting and reviewing) across language boundaries and historical periods. However, six researchers cannot cover every language across the 1710–1920 period. In order to make our data as rich as possible, we will be inviting researchers from outside our research team to contribute missing data. We will develop an online workflow, based as far as possible on existing crowd-sourcing initiatives, for managing this community-sourced content. The workflow will enable the community-sourced data to be reviewed by the research team before it is added to the *Agents of Change* dataset. An important aspect of the workflow will be to ensure that such contributors are properly credited for their work.

Alongside the collection of our research data, we are also requesting retrospective assignment of International Standard Serial Numbers (ISSN), a unique identification code for serial publications, for the periodicals which do not currently have one. These ISSNs will be used as stable identifiers for the periodicals we store within our database. Similarly, we are working on establishing authority records via the Virtual International Authority File (VIAF) for those women editors that do not have one yet. We will use use their VIAF IDs as stable identifiers within our NodeGoat dataset.

Our ultimate goal is to create a Virtual Research Environment (VRE), as an essential tool for establishing the research field of European Periodical Studies. The VRE will bring together primary sources and secondary literature, as well as the original scholarship that is produced as a result of the research data that we collect. Additionally, we would like the VRE to enable researchers outside the project team to contribute to the field, which we hope, step-by-step, will become a research community for European Periodical Studies.

We want to ensure that *Agents of Change* becomes a sustainable research tool beyond the end of the project funding by working closely with the local digital humanities centre. The Ghent Centre for Digital Humanities (GhentCDH) is an interdisciplinary centre facilitating digitally-enabled research in the arts, humanities and social sciences at Ghent University and beyond. GhentCDH also plays an active role in the coordination of Belgium's participation in DARIAH, the Digital Research Infrastructure for the Arts and Humanities. Within the framework of DARIAH, Ghent University, along with the universities of Antwerp and Leuven, has received startup funding from the Research Foundation Flanders, to develop a Virtual Research Environment Service Infrastructure (VRE-SI).

The VRE-SI is being developed by focussing on the infrastructural needs of existing humanities research projects in Flanders and Belgium that have a 'digital focus'. Now that the first year of the DARIAH-Flanders project has drawn to an end, the project team have gained a better understanding of how DARIAH Partner Institutions can sustainably support digital scholarship in the humanities. For example, at Ghent University, it has been identified that the establishment of a *digital humanities expert team* including humanities researchers, library staff, IT professionals and digital humanities experts would help to institutionally embed digital humanities research support.

The role of this interdisciplinary team is to both support the realisation of the digital humanities aspects of existing humanities research projects as well as providing advice and guidance in the development of new project proposals. To date, it has been identified that a missing element in the existing service provision is a *digital humanities scientific programmer*, whose role is to combine an understanding of the humanities research questions with the skills of an IT professional to realise the tools and services needed. It is possible to use the DARIAH funding to temporarily recruit such a member of staff and to demonstrate the value of such a post to the Faculty Management Team, with the view to, such a position being structurally funded by the university, in the medium to long-term.

Considering the curation and management of the research data both during the project funding and beyond, is a further crucial aspect of the project. As it is intended that *Agents of Change* will become a sustainable digital humanities research tool thriving beyond the fixed-term project funding, the establishment and implementation of a Data Management Plan (DMP) has been anticipated from the start of the project. The Faculty Library of Arts and Philosophy, as a result of their Arts and Humanities Research Data project, coordinated by their LibraryLab, is providing support to researchers in the faculty in the development of DMPs. Within DARIAH-BE, the intention is that every 'DARIAH pilot project', is strongly recommended to consider research data management from the outset. Finally, the GhentCDH is working closely with Lab1100 to explore how NodeGoat could be offered as a DARIAH-service. In the medium to long-term, the possibility of facilitating the development of an open source community around NodeGoat to further extend the environment for the needs of the digital humanities research, is being investigated.

The aim of this short paper is to firstly, to present some initial research results, based on the analysis of the data gathered by the *Agents of Change* on tracing the networks of intellectual exchange across temporal, geographic and linguistic borders through women editors and their periodicals. Secondly, this paper will demonstrate how working together with the local digital humanities centre and participating in DARIAH is helping to facilitate *Agents of Change* in becoming a sustainable digital humanities research tool that will thrive beyond the end of the fixed-term project funding.

Erlebter Raum im Rom der späten Republik - eine digitale Forschungsumgebung

Leif Scheuermann
leif.scheuermann@uni-graz.at
Alte Geschichte und Altertumskunde, Karl-Franzens Universität Graz, Austria

Klaus P. Jantke
klaus.jantke@idmt.fraunhofer.de
Fraunhofer-Institut für Digitale Medientechnologie Ilmenau, Erfurt, Germany

Walter Scheuermann
scheuermann@ike.uni-stuttgart.de
Institut für Kernenergetik und Energiesysteme, Universität Stuttgart, Germany

Thema dieses Beitrags ist der Aufbau und die Nutzung einer digitalen Forschungsumgebung zur Analyse und Visualisierung textueller, individueller und gruppenbasierter Raumwahrnehmung und Raumordnung der Stadt Rom im ersten vorchristlichen Jahrhundert. Neben einer grundlegenden kurzen Einführung in die historische bzw. raumwissenschaftliche Fragstellung sollen insbesondere der Umgang mit unscharfen bzw. unsicheren räumlichen Daten sowie der technischen Umsetzung einer digitalen Forschungsumgebung, in der unterschiedlichste Daten und Anwendungen miteinander kombiniert werden können und so ein genuin digitaler hermeneutischer Prozess ermöglicht und dokumentiert werden kann, im Fokus stehen.

Theoretische Grundlage

In den letzten Jahren haben die sich neu etablierenden Raumwissenschaften besonders für die Neuzeit ein Modell entwickelt, welches den historischen Raum von einem vermessbaren „euklidischen" oder auch einem positivistisch erfassten Natur- bzw. Kulturraum unterscheidet. Es konnte herausgestellt werden, dass Raum nicht die Summe und Relation der Objekte zueinander, sondern dass er erst im Aneignungsprozess durch die historischen Akteure generiert wird. Raum kann also nicht mehr von seinem Inhalt getrennt angesehen werden, sondern fungiert als dessen Ordnungsgefüge [PARNREITER 2013 S. 46] oder wie Ernst Cassirer es bereits 1930 programmatisch in seinem Vortrag „Mythischer, ästhetischer und theoretischer Raum" auf den Punkt gebracht hat, die Substantialität ist durch die Funktion, das „Was" durch ein „Wie" ersetzt worden [CASSIRER 2009/1931 S. 95]. In Bezug auf das Verhältnis von Geschichte und Raum bedeutet diese Öffnung des Raums vom Absolutum zum Betrachtungsgegenstand, a) dass die Vorstellung von Geschichte im Raum, bzw. der Raum als Behälter von Geschichte in eine Raum-

Geschichte umgewandelt werden muss, b) dass es nicht Eine Raum-Geschichte, sondern eine Pluralität von Raum-Geschichten gibt, welche vom Standpunkt bzw. der Lebenswelt des Betrachters abhängig sind und c) dass innerhalb dieser historischen Raum-Ordnungen Platz für Widersprüche, Leerstellen und Ungereimtheiten bestehen darf, ja muss. Denn in der Konstitution dieses „Wie“, dieser Raum-Ordnungen der Lebenswelt, spielt neben der sensorischen Wahrnehmung die individuelle Erfahrung des bzw. der Wahrnehmenden ebenso eine Rolle wie auch das kollektive Gedächtnis der jeweiligen Gesellschaft [LYNCH 1960/2001 S. 13]. Ziel einer historischen Raumwissenschaft muss es also sein, diese Pluralität von Raum-Ordnungen in ihrer vernetzten Struktur zu erfassen und auf ihre individuellen und gruppenspezifischen Anteile hin zu untersuchen. Konventionelle historische Karten können für die Analyse dieser Prozesse nur eine Grundlage bilden, auf welcher die mentalen Karten der historischen Akteure aufbauen, die wiederum die Basis für das Handeln Letzterer darstellen.

Historisches Szenario

Aus diesen grundlegenden Überlegungen zum Raum ergibt sich für eine Untersuchung der antiken Raumordnungen von Städten im Allgemeinen und der Stadt Rom im Speziellem als erstes Ziel der Aufbau der vermittelten Räume, aus den auf uns überkommen Quellen – in diesem Fall besonders den Texten. Zu Beginn müssen also, in dem hier vorgestellten Projekt vom lateinischen Autor Marcus Tullius Cicero (106-43v.Chr.) präsentierten Räume erfasst und dargestellt werden. Praktisch bedeutet dies zu fragen, welche räumlichen Objekte genannt werden, welcher Stellenwert ihnen zugemessen wird und wie sie zueinander in Verbindung gesetzt werden.

In einem zweiten Schritt muss ferner die Frage gestellt werden, welche individuellen Aneignungen und welche kollektiven Vorstellungen der Räume durch diese vermittelten Räume zutage treten und welchen Charakter diese besitzen. Handelt es sich um staatlich bzw. durch zentrale Akteure geplante, um alltägliche oder auch um symbolische Räume? Können diese klar voneinander getrennt werden oder überlagern bzw. bedingen sie sich und, sollte letzteres der Fall sein, auf welche Weise?

Die Quellenbasis für diese Analyse besteht in 1300 Nennungen von insg. 255 Raumobjekten der Stadt Rom in den Werken Ciceros, welche in einer relationalen Datenbank gesammelt und georeferenziert wurden. Ein zentrales Problem hierbei bestand in der Unschärfe der genannten Räume, da a) nicht alle Orte, die Cicero nennt, heute genau zu lokalisieren sind, b) die Nennungen, die der Autor macht, nicht immer exakt sind und c) manche Verortungen in der Zeit der Entstehung der Werke nicht mehr oder nur sehr vage bekannt waren (z.B. wenn sie sich auf die Zeit der Gründung Roms beziehen). Als Referenzdaten für eine räumliche Analyse wurden der archäologische Befund zur Stadt Rom in der späten Republik kartographisch aufgenommen sowie Nennungen bei für Cicero zeitgenössischen Autoren (Varro, Caesar, Sallust) und die historischen Daten zur Stadt Rom in der späten Republik.

Technische Umsetzung

Zentral für die Umsetzung eines solchen Ansatzes ist es, ein dynamisches kartographisches System aufzubauen, indem unterschiedlichste archäologische, philologische und historische Daten zusammengeführt und unter Nutzung verschiedenster Analyse- und Visualisierungswerkzeuge (kartographisch, netzwerkanalytisch, word-Clouds …) dynamisch miteinander in Beziehung gesetzt werden können. Es bedarf hierfür einer digitalen Forschungsumgebung, in der unterschiedliche Medieninhalte miteinander über streng definierte Programmierschnittstellen kombiniert werden.

Einen Schlüssel für die Verbindung unterschiedlicher digitaler Analyseverfahren bei freier Manipulierbarkeit der Daten bildet die Webble-Technologie [TANAKA 2003], eine Schnittstellentechnologie, mit der versucht wird, philosophische Konzepte der Wissensevolution in innovativen Werkzeugen des Wissensmanagements umzusetzen. Webbles erlauben Nutzern, vorhandene Wissensressourcen, welche als Medienobjekte gekapselt – „gewrapped“ – sind, weiterzuverarbeiten und zu distribuieren. Benutzer können einzelne Medienobjekte durch direkte Manipulation, wie „drag“, „drop“, „copy“, „paste“, miteinander zu neuen Objekten kombinieren, ohne Programmierkenntnisse zu besitzen. Aus technologischer Sicht handelt es sich bei der Webble-Technologie um eine Middleware, die insbesondere für Web-basierte Anwendungen geeignet ist und die intuitive Verbindung von nahezu beliebigen Funktionen und Dienstleistungen erlaubt. Dies gilt für die Verbindung von Methoden der qualifizierenden Datenanalyse mit denen der Bildverarbeitung gleichermaßen wie etwa der Verbindung von GIS-basierten Visualisierungssystemen mit digitalen Methoden der Netzwerkanalyse. Die Webble-Technologie erlaubt die intuitive Kombination von Medienobjekten, die im Ergebnisse neue Sichten oder gar Einsichten repräsentieren können.

Bibliographie

Parnreiter, C.(2013). *Stadtgeographie*. In: Harald Mieg, Christoph Heyl: Stadt. Ein interdisziplinäres Handbuch. Stuttgart, Weimar: Verlag J.B. Metzler, pp. 46-63.

Cassierer, E. (1931/2009). *Mythischer, ästhetischer und theoretischer Raum*. In: Wolfgang ORTH und Michael KROIS (Hg.): Ernst Cassierer. Symbol, Technik, Sprache. Aufsätze aus den Jahren 1927-1933. Hamburg: Felix Meiner Verlag (Philosophische Bibliothek 372), pp. 93-110.

Tanaka, Y. (2003). *Meme media and meme market architectures. Knowledge media for editing, distributing, and managing intellectual resources*. Piscataway, NJ, Hoboken, NJ: IEEE Press; Wiley-Interscience.

Linking Qualitative and Quantitative Research by Thin Descriptions through Semantic Graphs. Experiments in Apparatus Design with Semantic CorA.

Christoph Schindler
schindler@dipf.de
German Institute for International Educational Research (DIPF)

Basil Ell
basil.ell@kit.edu
Karlsruhe Institute of Technology (KIT)

Introduction

The use of Semantic Web Technologies in Digital Humanities projects has increased over the years, but only recently annotation tools and research environments have started to create and use semantic graphs (e.g. Pundit[1], Grassi et al. 2013, or CWRC[2], Rockwell et al., 2012). While the sharing of annotations and data into a formalized Web of Data is a central part of these projects, in this paper we would like to focus on capacities for qualitative research by using semantic graphs and linking these with quantitative data. Qualitative research approaches have thus far not been positioned at the core of Digital Humanities (Drucker, 2012). However, recent discussions about close and distant reading e.g., (Moretti, 2013) point out a manifestation of a strict opposition between qualitative and quantitative research on the level of data. We would like to reconfigure this separation between qualitative and quantitative data by taking into account its epistemological practices and apparatuses. Qualitative and quantitative data can be regarded as "thin descriptions". Both need an interpretative act to become "thick descriptions" (Love, 2013). This involves a broad range of research capacities, whereby a semantic graph addresses parts of it and creates a link between both paradigms by connecting the different data. This main argument of the paper will be demonstrated through examples from ongoing Digital Humanities research projects based on the semantic research environment Semantic CorA[3], where we experimented with epistemological apparatuses which were developed through participatory design approaches and collaborative ontology engineering.

Re-Configuring the apparatus of qualitative and quantitative research

The recent focus in Digital Humanities on epistemological apparatuses, including their materiality, performativity and relation to theory, offers the possibility to sharpen the respective analytical concepts for data and research apparatuses. Ramsey and Rockwell (2012) describe tools as a "telescope for the mind" and offer a materialistic epistemology. Concerning the opposition of qualitative and quantitative research, Manovich (2011) describes the potentials of big data by contrasting quantitative methods (i.e., statistical, mathematical, computational) with qualitative methods (i.e., as used in History, Literature Studies, Anthropology, qualitative Social Sciences and Psychology) and the different kinds of underlying data. While quantitative approaches commonly rely on surface data, qualitative data are described as deep data. Manovich posits equivalent epistemological depth of both kinds of data but hints to the different scale of contact points with the object of interest.

Venturini and Latour (2010) point to the micro/macro distinction in Social Sciences, which corresponds to the qualitative/quantitative separation at the methodological level. The new capacity of digital, computerized methods, they point out, are quali-quantitative methods, which do not rely on the opposition of statistical analysis and ethnographic observation. Love (2013) concretizes this link between qualitative and quantitative approaches as well and takes into account the epistemological apparatus by comparing Literature Studies and ethnographic research. Instead of placing thin description in opposition to "thick description" (Geertz, 1973), Love (2013: 403) argues for the significance of thin description, which she considers as an integral part of thick description, and demands for a reflective engagement with the full range of empirical methods. In this sense the epistemological apparatuses are per se boundary-making practices, which "enact what matters and what is excluded from mattering" (Barad, 2007: 148) and demand an accountability of this material-discursive practice of epistemological apparatus design.

Designing Semantic CorA

The realization of Sematic CorA was driven by a participatory design and agile development approach of both the software components as well as the ontology. A main goal was to enable classical qualitative researchers to realize their research by using the potentials of a semantic graph. The design followed the research practice though an evolutionary process, which was initiated by an analysis of needs and requirements (i.e., site visit, artefact analysis, interviews) and was followed by nearly weekly meetings with requirement articulations and prototype testings. Semantic MediaWiki[4] was used as a technological platform, which was extended with newly developed tools (extensions) to address the needed research capacities.

Linking qualitative and quantitative Research with Semantic CorA

Openness and flexibility: In qualitative research, the openness of the research object and the flexibility to adjust the knowledge base play a central role in grasping the complexity of the phenomenon of interest (Bauer, Aarts, 2007). Using a semantic graph as an epistemological tool offers the possibility to create new entities, add properties or relate these to other entities. In this way, a network is created step by step based on the research material, which can be extended and re-arranged in the research process. So an all-encompassing fixed schema is not a precondition when starting a research project. The semantic graph consists of pages (representing entities), which are described by properties or links. Figure 1 demonstrates the schema of the semantic graph for exploring historical educational lexica (ranging from 1774 to 1945). Integrated bibliographic data (e.g., lexicon, article, author) and digital images of the lexica from a digital library[5] build the data basis for the graph, whereby the syntax supported by the SMW platform and data edit forms offer to enrich the graph. For example, a researcher establishes the relation `Is About´ between the article and a person and adds further properties (date of birth/death, gender, nationality, domain of mattering). In this respect, a thin description has been created, indicating a network of the phenomenon with more than 6 million values and 100,000 pages.[6]

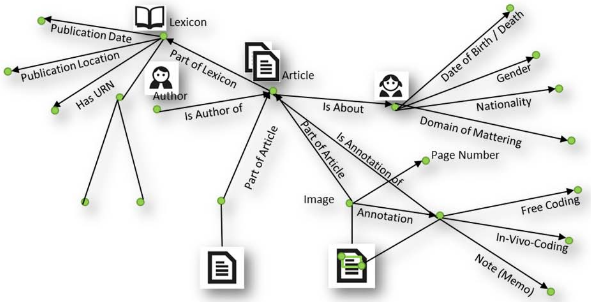
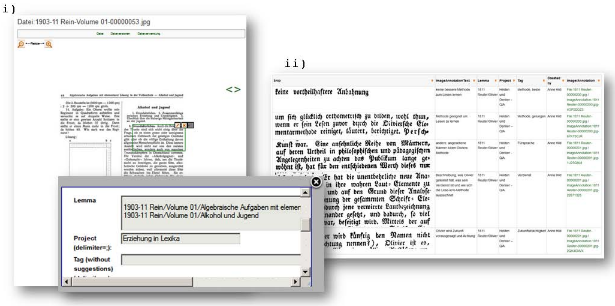


Figure 1: Schema of the semantic graph for exploring educational lexica

Balancing particularities and formalizations: In qualitative research it is necessary to balance formalization with respect to the particularities of the phenomenon of interest and the research material. Star describes this balance as achieved through facetted classifications in the methodological approach of the grounded theory (Star, 1998: 227). As previously described, a semantic graph offers the possibility to create a network, which calls for formalization and sets the boundaries of the phenomenon. To enable a more `fuzzy´ or qualitative thin description, each node and property of the graph can be described in an unformalized way using the classical text tools of a wiki. Additionally, an open approach for qualitative content analysis is followed to enable annotations of the text (Figure 2). In this way, the annotation is connected to the article on the basis of the semantic graph (Figure 1), providing links between the annotations of the qualitative content analysis and the further network of the semantic graph. This constitutes a thin description based on qualitative data (e.g., through close reading and annotations) and quantitative data (e.g., bibliographic data), where an interpretative act is needed for the description to become a thick description.



Datei:1903-11 Rein-Volume 01-00000053.jpg

Lemma: 1903-11 Rein/Volume 01/Algebraische Aufgaben mit elemer
1903-11 Rein/Volume 01/Alkohol und Jugend

Project (delimiter=;): Erziehung in Lexika

Tag (without suggestions):

Snip	ImageAnnotationText	Lemma	Project	Tag	Created by	ImageAnnotation
keine vortheilhaftere Anbahnung	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
um ſich glücklich orthometriſch zu bilden, wohl thun, wenn er ſein Leſen zuvor durch die Olivierſche Elementarmethode reiniget, läutert, berichtiget. Perſch-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Kunſt war. Eine anſehnliche Reihe von Männern, auf deren Urtheil in philoſophiſchen und pädagogiſchen Angelegenheiten zu achten das Publikum lange gewöhnt iſt, hat für den entſchiedenen Werth dieſes nur	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Er hat die unentbehrliche neue Ana- in ihre wahren Laut- Elemente zu und auf den Grund dieſer Analyſe nung der geſammten Schrift- Ele- durch jene verwirrte Lautbezeichnung nander geſetzt, und dadurch, ſo viel war, beſeitigt wird. Mittels der auf	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
er wird künftig den Namen nicht chtung nennen?), Olivier iſt es,	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Figure 2: **Annotation tool with data edit form** (i) and **a**ggregated annotations in a dynamic table (ii)

Ongoing data analysis and graph creation: Another main aspect of qualitative research is the ongoing iteration between grasping the research material and creating new connections or qualities. This intensive work with research data can be done by comparing, following associations, close reading, or – as previously described – by distant reading of aggregated or related elements. While using a semantic graph as a network of the phenomenon, two different examples can be demonstrated which use the linkage between qualitative and quantitative data: Semantic browsing and querying the network. To represent relevant parts of the graph, aggregations and inferences are created for main entities (e.g. 1,200 aggregated descriptions of persons) to thickening the research data and browsing through the network (Figure 3, i). Additionally, a query tool enables to query and aggregate the semantic graph (Figure 3, ii).

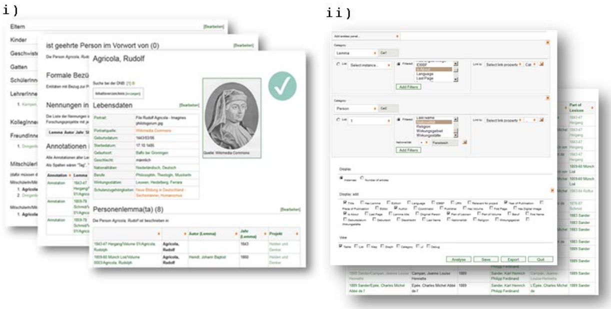


Figure 3: Semantic Browsing through the aggregated information on page about person (i) and analysis tool with results in dynamic table (ii).

Discussion and outlook

This paper demonstrates the possibility of linking qualitative and quantitative data by using semantic graph technologies to create thin descriptions. Therefore, the advantages of the interpretative act of thick descriptions are considered by allowing for ongoing iterations of analyzing the research material and creating the semantic graph in a formalized and unformalized way (enrichments, annotations, text descriptions). With semantic browsing, aggregations of information, annotation and querying the semantic graph, aspects of close and distant reading are addressed, thus offering new techniques for grasping the research material for qualitative research.

For the Digital Humanities, the focus on mattering of apparatuses offers the possibility to open the design space for digital tools to the diversity of epistemological practices in Humanities. Thereby, an engagement with the diversity of the Humanities comes to the front, enhancing the accountability of boundaries and possibilities of epistemological apparatuses in Digital Humanities.

Acknowledgements

The authors would like to thank the research group around Semantic CorA, especially Marc Rittberger, Lia Veja, Kendra Sticht, Anne Hild, and Anna Stisser. The initial realization of the research environment Semantic CorA was supported by the German Research Foundation (DFG) and its further development is supported in the context of CEDIFOR by the eHumanities program of the German Federal Ministry of Education and Research (BMBF) no. 01UG1416C.

Bibliography

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Bauer, M. W. and Aarts, B. (2007). Corpus Construction: a Principle for Qualitative Data Collection. In Bauer, M. W. and Gaskell, G. (Eds.), *Qualitative Researching with Text, Image and Sound: A Practical Handbook*. London: Sage, pp. 19–37.

Drucker, J. (2012). Humanistic theory and digital scholarship. *Debates in the Digital Humanities*, pp. 85–95.

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays*. Basic books.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S. and Piazza, F. (2013). Pundit: augmenting web contents with semantics. *Literary and Linguistic Computing*: fqt060 doi:10.1093/llc/fqt060.

Love, H. (2013). Close Reading and Thin Description. *Public Culture*, **25**(371): 401–34.

Manovich, L. (2011). Trending: the promises and the challenges of big social data. *Debates in the Digital Humanities*, pp. 460–75.

Moretti, F. (2013). *Distant Reading*. Verso Books (accessed 31 October 2015).

Ramsey, S. and Rockwell, G. (2012). Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In Gold, M. K. (Ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 75–84.

Rockwell, G., Brown, S., Chartrand, J. and Hesemeier, S. (2012). CWRC-Writer: An In-Browser XML Editor. *Poster Presented at Digital Humanities*.

Star, S. L. (1998). Grounded Classification: Grounded Theory and Faceted Classification. *Library Trends*, **47**(2): 218–32.

Venturini, T. and Latour, B. (2010). The social fabric: Digital traces and quali-quantitative methods. *Proceedings of Future En Seine*, pp. 30–15.

Notes

[1] http://thepund.it

[2] http://www.cwrc.ca

[3] http://semantic-cora.org

[4] http://semantic-mediawiki.org/

[5] http://bbf.dipf.de/digital-bbf/spo

[6] Additionally, the technological platform itself offers the possibility to link elements of the graph to established Semantic Web vocabularies (i.e., BIBO or DC); this has been performed for a large part of the data.

Making Sense of Illustrated Handwritten Archives

Lambert Schomaker
l.r.b.schomaker@rug.nl
ALICE, University of Groningen, The Netherlands

Andreas Weber
a.weber@utwente.nl
STePS, University of Twente, The Netherlands

Michiel Thijssen
thijssen@brill.com
BRILL, The Netherlands

Maarten Heerlien
maarten.heerlien@naturalis.nl
Naturalis Biodiversity Center, The Netherlands

Aske Plaat
aske.plaat@gmail.com
LIACS, Leiden University, The Netherlands

Siegfried Nijssen
s.nijssen@liacs.leidenuniv.nl
LIACS, Leiden University, The Netherlands

Fons Verbeek
f.j.verbeek@liacs.leidenuniv.nl
LIACS, Leiden University, The Netherlands

Michael Lew
lewmsk@gmail.com
LIACS, Leiden University, The Netherlands

Eulalia Gasso Miracle
Eulalia.GassoMiracle@naturalis.nl
Naturalis Biodiversity Center, The Netherlands

Katy Wolstencroft
k.j.wolstencroft@liacs.leidenuniv.nl
LIACS, Leiden University, The Netherlands

Ernest Suyver
suyver@brill.com
BRILL, The Netherlands

Bart Verheij
Bart.Verheij@rug.nl
ALICE, University of Groningen, The Netherlands

Marco Wiering
m.a.wiering@rug.nl
ALICE, University of Groningen, The Netherlands

Rene Dekker
Rene.Dekker@naturalis.nl
Naturalis Biodiversity Center, The Netherlands

Joost Kok
joost.n.kok@gmail.com
LIACS, Leiden University, The Netherlands

Lissa Roberts
l.l.roberts@utwente.nl
STePS, University of Twente, The Netherlands

Jaap Van den Herik
jaapvandenherik@gmail.com
LCDS, Leiden University, The Netherlands

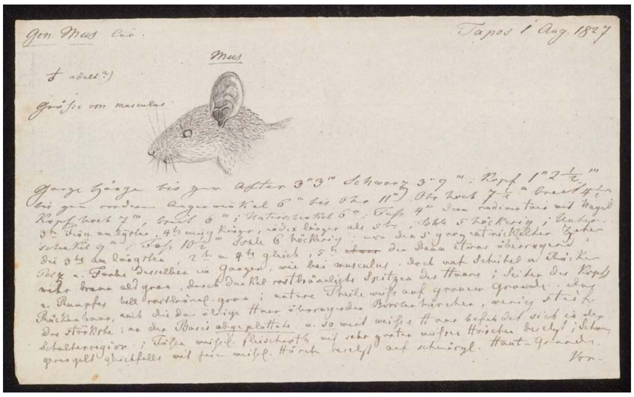

Gen. Mus Cuv.
Tapos 1 Aug. 1827
Mus
Grösse von musculus

Figure 1. Page from a bundle of field notes, describing and depicting a mouse species. Source: Naturalis Biodiversity Center, Archief van de Natuurkundige Commissie voor Nederlands-Indië. Copyright: Public Domain Mark 1.0

The MONK system uses shape-based feature vector methods that have very few assumptions concerning the content or style of the material. It avoids the traditional OCR approach (optical character recognition) which assumes that individual characters are essentially legible. That assumption only holds for a tiny fraction of handwritten material and a limited number of scripts. The only assumptions MONK makes are that pictorial and textual segments are separated by white spaces; and that the layout, of underlining, etc. in a specific document, is consistent throughout the document. In MONK, classification methods are used that allow for a fast bootstrap from single example instances (nearest-neighbor search) (Gast et al., 2013). With larger numbers of labeled examples, models can be computed, varying from nearest-centroids to support-vector machines and neural networks in a continuous learning process (Krizhevsky et al., 2012; Liu et al, 2015; Guo, in press). A challenging topic from the technical point of view is the relation between existing semantic knowledge (ontologies) and the statistically inferred semantics using Google's *word2vec* and current deep-learning neural networks. Can the underlying structure and style in a collection of a common and realistic size be detected by such algorithms? Can the proposed enrichment system profit from generally available text corpora? The processing power required by the proposed architecture is substantial. For this project, algorithmic optimization of the image processing and recognition system is necessary in order to create the necessary speed and flexibility of the system for use by non-technical end users. In order to tackle this challenge the consortium will make use of the combined knowledge and expertise of ALICE in Groningen, and the Leiden Institute of Advanced Computer Science (LIACS), where multiple supercomputers and high performance computing experts are present.

Because of its visual approach, MONK can handle the diversity of material that we encounter in our use case and in historical collections in general: text, drawings, and images. MONK also does not require a language model nor fully transcribed samples to quickly assess the contents of an archive page. The human-in-the-loop approach of MONK is currently 'label' oriented, but will be enhanced by providing the user and the system with ontologies for bootstrapping the learning process. The system will understand handwritten corpora to such an extent that the visual and textual content on individual pages is categorized, determined and networked to other pages in the archive and external sources. To construct training examples for MONK, biologists and historians of science will manually label documents to the machine learning software by means of a human in the loop approach. In addition, a crowdsourcing approach will be used to further expand this corpus of examples. Our consortium will here build on the expertise of ALICE and Naturalis Biodiversity Center, the Leiden-based National Museum of Natural History. Eventually, the computer-assisted recognition of words and visual information on a page will thus allow users to search, filter and group arbitrary archive items and enables connections with external databases. Last but not least, MONK lays the groundwork for full transcription of any handwritten-illustrated archival collection.

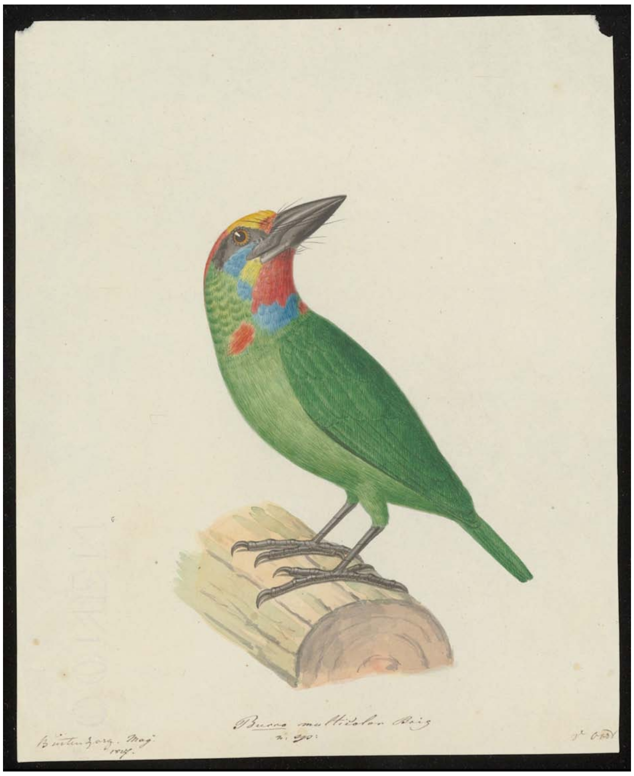


Figure 2. Drawing of Burro multicolor created in Buitenzorg, Java in 1827 by Pieter van Oort. Source: Naturalis Biodiversity Center, Archief van de Natuurkundige Commissie voor Nederlands-Indië. Copyright: Public Domain Mark 1.0

The central use case of our research project is the collection of the *Natuurkundige Commissie voor Nederlandsch-Indië* (hereinafter NC). It is one of the top-collections of Naturalis. From 1820 to 1850, the NC charted the natural and economic state of the Dutch East Indies and returned a wealth of scientific data and specimens which are now stored in archives in the Netherlands and Indonesia. The collection comprises thousands of handwritten notes and drawings and tens of thousands biological and geological specimens. While these archival items have all been digitized, the individual pages in the notebooks, diaries and reports are not catalogued nor labeled separately. Many of the field notes combine different textual and visual elements on one page. Our short paper presentation is based on the processing of an initial set of understudied handwritten field notes which we carried out in early 2016. By doing so, we will demonstrate the efficiency of the MONK system and our approach.

Owing to the different 'hands' and languages used in the documents, links across handwritten field records and notes, drawings and specimens cannot be made in an efficient way. Our corpus contains material from at least seventeen different writers and the used languages range from German *(Kurrentschrift)* to Latin, French, and Dutch. The labels of related historic specimens only provide very general information on collection localities and collectors. Hence, the typical use case of a scholar wishing to retrieve information on a certain species, person, drawing, or collecting locality is limited. Owing to its sheer dimension and its weak structure, it is impractical to disclose and network this archive manually. Its current inaccessibility hampers research into Southeast Asian natural history and the history of (scientific) knowledge production. Knowledge extracted from the documents will be structured and served as Linked Open Data. This will allow interlinking of content and also enable interoperability with other cultural heritage resources, for example, the physical specimens obtained during expeditions, or other historically significant data collections from the same area.

The multi-layered character of the material makes it a perfect use case for developing a technologically advanced and usability engineered digital environment for interpreting and connecting illustrated-handwritten collections. In our consortium data scientists from the Universities in Leiden and Groningen work closely together historians of science from the University of Twente and taxonomy experts from Naturalis. Fueled by an investment from BRILL publishers in a national funding scheme, this project will not only result in the disclosure of the NC archive, but will also enable the integrated study of underexplored scientific manuscript collections and archives in general.

Bibliography

Zant, T. van der, Schomaker, L. and Haak, K. (2008). Handwritten-Word Spotting Using Biologically Inspired Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(11): 1945–57.

Oosten, J.-P. van and Schomaker, L. (2014a). Separability versus prototypicality in handwritten word-image retrieval. *Pattern Recognition*, **47**(3): 1031–38. (Handwriting Recognition and Other PR Applications)

Oosten, J.-P. van and Schomaker, L. (2014b). A Reevaluation and Benchmark of Hidden Markov Models. *2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. pp. 531–36.

Gast, E., Oerlemans, A. and Lew, M. S. (2013). Very large scale nearest neighbor search: ideas, strategies and challenges. *International Journal of Multimedia Information Retrieval*, **2**(4): 229–41.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. pp. 1097–105.

Liu, Y., Guo, Y., Wu, S. and Lew, M. S. (2015). DeepIndex for Accurate and Efficient Image Retrieval. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. (ICMR '15). New York, NY, USA: ACM, pp. 43–50.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M. S. (2015). Deep learning for visual understanding: A review. *Neurocomputing*. (Available online 26 November 2015).

Digital History "From Below": a Call to Action

Anelise Hanson Shrout
anshrout@davidson.edu
Davidson College, United States of America

Humanists – inclusive of digital humanists – are preoccupied with telling stories. Some of our most interesting subjects, however, have left only the barest of marks on historical records. Their stories are among the most captivating, but also some of the most difficult to access. This paper knits together recent trends in digital humanities practices that have helped us to elevate unrepresented voices with a discussion of how to elevate the marginalized within the DH community. It showcases select projects that undermine archival silences.[1] It then argues that digital humanities practitioners should add these theories to the collection of tools currently used to forward social justice projects in DH spaces.

Elevating the Archivally Silenced

Various methodologies have been adopted to address the problem of how to tell stories about people who left behind few records. In the 1970s and 1980s, practitioners of "history from below" worked to elevate narratives about "people with no history," by chronicling the everyday lives of peasants and non-elites. At the same time, practitioners of the "new social history" turned to cliometrics – and adopted methods that would be familiar to those who work with "big data" today – to highlight trends about marginalized peoples from historical data like censuses, probate records and financial documents.

There have been various resurgences and developments in these methods in the intervening four decades. These include practices of reading archives "against the grain" to get at the unstated assumptions that historical actors made about those they held power over. They also include theoretical approaches that advocate the reading of silences to understand those whose voices were intentionally obscured by official recorders and gatekeepers.

Marginalizations Within DH

Questions about whose voices are elevated and whose are silenced have also long been a theme in DH scholarship and discourse. These questions seek to unpack the ways in which DH as a field is exclusionary. This former is a much (though still not enough) referenced problem in panels at former DH conferences, which have asked how DH research can address (and remedy) social problems.

Digital humanities scholarship has also begun to address problems of access within the broader DH community, and the barriers erected to women and people of color in particular. For example, Adeline Koh has argued that we need to examine the ways in which DH publics are constituted, in order to better understand the creation of "limits of the discourse that defines the idea of a digital humanities 'citizen.'" Similarly, Tara McPherson has argued that we must see the evolution of DH as a field shaped by structural inequalities – of race, class and gender – which accompanied the rise of computation technologies.

A Knitted View

These are much needed interventions, and help us to understand the evolution of our field as one in which certain groups have been marginalized and others have been centered. These conversations also mirror methodological debates within history about whose voices to elevate, and under what circumstances. This paper complements extant work by arguing that theoretical interventions concerning current structural inequalities must be brought to bear on the past, and that digital methodologies are ideally suited to elevating subsumed voices in the present. It further demonstrates that these projects, the theories that underlie them, and current work to make DH more equable should be read together to further the practice of digital history and humanities "from below.

Bibliography

Bastian, J. (2003). *Owning Memory: How a Caribbean Community Lost Its Archives and Found Its History*. Westport, Conn.: Libraries Unlimited.

Bhattacharya, S. (1983). History from Below. *Social Scientist*, 3(20).

Farge, A. (2015). *The Allure of the Archives*, New Haven: Yale University Press

Fuentes, M. (2010). Power and Historical Figuring: Rachael Pringle Polgreen's Troubled Archive. *Gender and History*, **22**(3): 564–84.

Gallman, R. (1977). Some Notes on the New Social History. *The Journal of Economic History*, **37**(1): 3–12.

Koh, A. (2014). Niceness, Building, and Opening the Genealogy of the Digital Humanities: Beyond the Social Contract of Humanities Computing. *Differences*, **25**(1): 93–106. doi:10.1215/10407391-2420015.

McPherson, T. (2012). Why Are the Digital Humanities So White? Or Thinking the Histories of Race and Computation. In Gold, M. (Ed.), *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press.

Trouillot, M. (1995). *Silencing the Past: Power and the Production of History*. Beacon Press.

Notes

[1] These might include work like Ben Schmidt's, elaboration upon late twentieth-century cliometrics and use of "big data" methods to explore historical sources (http://benschmidt.org/projects/digital-humanities-research/); maps like Vincent Brown's "Slave Revolt in Jamaica" which uses sources produced

by slaveholders to argue for the agency and tactical prowess of enslaved people (http://revolt.axismaps.com/map/); and Michelle Moravec's use of metadata to "unghost" lesbian women in the past (http://michellemoravec.com/).

Change, Transition and Governance: Lessons from a Long-Term, Large Scale DH Collaboration

Lynne Siemens
siemensl@uvic.ca
University of Victoria, Canada

Digital Humanities (DH) is becoming increasingly a collaborative community of practice, a move encouraged both by the scale, scope and complexity of projects (L. Siemens and Burr, 2013) and by granting agencies with programs such as Digging into Data (2013), the new Bilateral Digital Humanities Program between the National Endowment for the Humanities' Office of Digital Humanities and the Deutsche Forschungsgemeinschaft, Germany's research office (2014), Canada's Social Sciences and Humanities Research Council's Partnership Grants (2013), and many others. This trend is further supported through efforts such as Fair Cite (2012) and the Collaborators' Bill of Rights (Off the Tracks, 2011) to more fully recognize project contributions through multi-authorship citation practices, University of Virginia's Praxis Program which provides collaborative project experience for graduate students (The Praxis Program at the Scholars' Lab, 2011a, 2011b, 2012, nd), and individual DH project charters outlining guidelines for team work (Hjartarson, Fast, and Hasenbank, 2011; Ruecker and Radzikowska, 2007).

These are all exciting developments that will reap long-term contributions at the individual, project, and DH as a community of practice levels and beyond. At the same, more work is needed to understand how these teams function and the types of supports which are needed to coordinate the research, people and tasks to ensure successful outcomes at these levels (Dombrowski, 2013; Lyall, Bruce, Marsden, and Meagher, 2013). With this knowledge, DH teams will be better able to anticipate benefits and challenges associated with collaborations and develop processes to maximize benefits while minimizing associated challenges.

At present, much of the body of knowledge about academic team functioning and best practice guidelines has been developed through a reflection process at a project's completion (For example, see Bracken and Oughton, 2006; Bryan, Negretti, Christensen, and Stokes, 2002; Kishchuk, 2005; Lawrence, 2006; National Endowment for the Humanities Office of Digital Humanities, 2010). These reviews often focus on the actual research work accomplished with little discussion of associated processes that supported the work, communication patterns, and other factors that supported (or not) successful outcomes. As a result, some hard-earned lessons are forgotten or minimized through the passage of time, but might be captured if this type of reflection occurs during a project's life.

As a large project in terms of team membership, budget, scope, disciplinary perspectives, and project length and research integration, Implementing New Knowledge Environments (INKE) provides a unique perspective to explore the nature of collaboration (INKE, 2012a). INKE's primary research focus is the exploration of e-books and their potential from a variety of perspectives, including interface design, modeling and prototyping, user studies, and textual studies (INKE, 2012b). Further, this collaboration is examining the "understanding, creating, and evaluating research structures that will allow academic and non-academic (including industry partner) members of our research team to work together in ways that meet the needs of the research and development cycles of the entire INKE group" (R. G. Siemens, Siemens, Cunningham, Galey, Ruecker, and Warwick, 2012, p. 7)[1]. To that end, INKE has undergone yearly reflections on the nature of collaboration within the project with an objective to better understand the ways to support large-scale research collaborations as they unfold and communicate these lessons to other projects for consideration.

This paper contributes to our understanding of how research teams function by reviewing INKE's six years of experience in collaboration and with a view to articulate best practice guidelines (L. Siemens and INKE Research Group, 2010, 2012b, 2012c, 2012d, 2012e, 2013).

Members of the administrative team, researchers, graduate research assistants and others are asked about their experiences collaborating within INKE on an annual basis in order to understand the nature of collaboration and ways that it may change over a grant's long-term life. The interview questions focus on understanding the nature of collaboration, advantages and challenges associated with it within INKE's context. These interviews allow the researcher to explore topics more fully and deeply with probing and follow up questions while participants reflect on their own experiences and emphasis those issues which are important to them (Marshall and Rossman, 1999; McCracken, 1988; Newell and Swan, 2000; Rubin and Rubin, 1995). This paper focuses on a summary of the first 5 years of the grant-funded work.

At the time of writing this proposal, final data analysis is being completed, but clear patterns are emerging and, after final analysis, these will form the basis of my presentation.

While the grant application suggested a stable team of active researchers and partners, the reality has been very different. Due to a variety of reasons (L. Siemens and

INKE Research Group, 2012a, 2013), change and transition have been constant within INKE, which has led to sub-research group re-organizations and the creation of new ones. Concurrently, new researchers, administrative leads and partners joined the team. And as is always the case when working with student research assistants and post-doctoral fellows, sub-research areas were continually recruiting and training new ones as others moved on to other opportunities.

Grounding this change has been several constants that have ensured the research has been able to continue effectively and efficiently despite the transitions. First, the governance documents outlined clear articulation of roles and responsibilities which became especially important for ensuring that new researchers, partners, and administrative understood and enacted the nature of collaboration and accountability within INKE. At the same time, these documents provided guidance for ensuring that processes for change, transition, planning and reporting, and decision making were considered and thoughtful while remaining responsive to changing circumstances (L. Siemens and INKE Research Group, 2012c). The use of basecamp, an online project space, and an updated project website further reinforced these processes by providing an ongoing repository for messages, documents, data, and publications, all important knowledge for current and new researchers and partners. Second, multiple communication channels, such as annual birds-of-a-feather gatherings, attendance at other conferences, both formal and in-formal face-to-face administrative and sub-research area group meetings, conference calls, and online project spaces, ensured that team members met on a regular basis to exchange information about and participate in research projects with the other sub-research areas, ensuring highly collaborative work.

As INKE nears the end of its 7 years of funded research, the team is anticipating both winding down this focused work on e-books and exploring next research steps, building from its successes (INKE, 2014). While the research has been intellectually challenging and not without its administrative issues, INKE team members report positive experiences in terms of the collaboration, associated outcomes, and connections to the DH and traditional humanities communities. As measures of this positive spirit and connections, they are extending their collaboration into associated research areas as well as contemplating another large-scale research project.

This research will make several contributions to the knowledge base about the nature of collaboration within the DH community of practice. First, this research contributes to efforts to make work patterns and relationships more explicit and understand those factors that tend to predispose them to success, and perhaps, more importantly, to avoid those which may lead to problematic interactions. Already, lessons from INKE are informing other projects' collaborations (Nowviskie, 2011; The Praxis Program at the Scholars' Lab, nd). Second, INKE's experience demonstrates that these types of team research projects require skills not typically taught in graduate school, including project management and collaboration within a targeted and integrative research environment, which differs from that of curiosity-based one. This reinforces the call to enlarge graduate training beyond purely disciplinary to these larger collaborative skills to ensure that students are prepared for both academic and alternative academic posts (Berman, 2011; Carr, 2012; Leon, 2011; Nowviskie, 2010; Powell, Bouchard, Dalgleish, Keenan, McLeod, and Thomson, 2013; L. Siemens, 2013; Spiro, 2010).

Bibliography

Berman, R. A. (2011). Reforming Doctoral Programs: The Sooner, the Better. Retrieved from http://www.mla.org/blog?topic=143

Bracken, L. J., and Oughton, E. A. (2006). 'What do you mean?' The importance of language in developing interdisciplinary research. *Transactions of the Institute of British Geographers, 31*(3), 371-382.

Bryan, L., Negretti, M., Christensen, F. B., and Stokes, S. (2002). Processing the process: One research team's experience of a collaborative research project *Contemporary Family Therapy, 24*(2), 333-353.

Carr, G. (2012). Graduate students need preparation for life outside the university, November 1, 2012, from http://m.theglobeandmail.com/news/national/graduate-students-need-preparation-for-life-outside-university/article4699319/?service=mobile

Digging into Data Challenge. (2013). Digging into data challenge, October 28, 2013, from http://www.diggingintodata.org

Dombrowski, Q. (2013). *What ever happened to Project Bamboo?* Paper presented at the DH 2013, Lincoln, Nebraska. http://dh2013.unl.edu/abstracts/ab-287.html

Fair Cite. (2012). Fair cite: Towards a fairer culture of citation in academia Retrieved March 5, 2012, from http://faircite.wordpress.com/

Hjartarson, P., Fast, K., and Hasenbank, A. (2011). *Modelling collaboration in digital humanities scholarship: Foundational concepts of an EMIC UA project charter*. Paper presented at the Space/Place/Play, Toronto, ON. http://www.cwrc.ca/events/final-ryerson-conference-program/

INKE. (2012a). About, October 29, 2012, from http://inke.uvic.ca/projects/about/

INKE. (2012b). Publications, August 28, 2013, from http://inke.ca/projects/publications/

INKE. (2014). Future directions, November 3, 2014, from http://inke.ca/projects/future-directions/

Kishchuk, N. (2005). Performance report: SSHRC's major collaborative research initiatives (MCRI) program. Ottawa, ON: SSHRC.

Lawrence, K. A. (2006). Walking the tightrope: The balancing acts of a large e-research project. *Computer Supported Cooperative Work: The Journal of Collaborative Computing, 15*(4), 385-411.

Leon, S. M. (2011). Project management for humanists: Preparing future primary investigators Retrieved June 24, 2011, from http://mediacommons.futureofthebook.org/alt-ac/pieces/project-management-humanists

Lyall, C., Bruce, A., Marsden, W., and Meagher, L. (2013). The role of funding agencies in creating interdisciplinary knowledge. *Science and Public Policy, 40*(1), 62-71.

Marshall, C., and Rossman, G. B. (1999). *Designing qualitative research* (3rd ed.). Thousand Oaks, CA: SAGE.

McCracken, G. (1988). *The long interview* (Vol. 13). Newbury Park, California: SAGE Publications.

National Endowment for the Humanities Office of Digital Humanities. (2010). Summary findings of NEH digital humanities start-up grants (2007-2010). Washington, D.C.: National Endowment for the Humanities,.

Newell, S., and Swan, J. (2000). Trust and inter-organizational networking. *Human Relations, 53*(10), 1287-1328.

Nowviskie, B. (2010). #alt-ac: Alternate academic careers for humanities scholars. Retrieved from http://nowviskie.org/2010/alt-ac/

Nowviskie, B. (2011). Where credit is due: Preconditions for the evaluation of collaborative digital scholarship. *Profession, 13*, 169–181.

Off the Tracks. (2011). Recommendations Retrieved May 26, 2011, from http://mith.umd.edu/offthetracks/recommendations/

Office of Digital Humanities. (2014). NHE/DFG 2014 bilateral digital humanities program, October 31, 3014, from http://www.neh.gov/divisions/odh/grant-news/nehdfg-2014-bilateral-digital-humanities-program

Powell, D., Bouchard, M., Dalgleish, M., Keenan, A., McLeod, A., and Thomson, T. (2013). Conversation, collaboration, credit: The graduate researcher in the digital scholarly environment. *Digital Studies/Le champ numerique, 4*.

Rubin, H. J., and Rubin, I. S. (1995). *Qualitative interviewing: The art of hearing data*. Thousand Oaks, CA: SAGE Publications.

Ruecker, S., and Radzikowska, M. (2007). *The iterative design of a project charter for interdisciplinary research*. Paper presented at the DIS 2007, Cape Town, South Africa.

Siemens, L. (2013). *Meta-methodologies and the DH methodological commons: Potential contribution of management and entrepreneurship to DH skill development*. Paper presented at the DH 2013, Lincoln, Nebraska.

Siemens, L., and Burr, E. (2013). A trip around the world: Accommodating geographical, linguistic and cultural diversity in academic research teams. *Linguistic and Literary Computing, 28*(2), 331-343.

Siemens, L., and INKE Research Group. (2010). *The e-paper anniversary: Lessons from the first year of INKE*. Paper presented at the SDH/SEMI 2010, Montreal, Quebec.

Siemens, L., and INKE Research Group. (2012a). Firing on all cylinders: Progress and transition in INKE's year 2. *Scholarly and Research Communication, 3*(4), 1-16.

Siemens, L., and INKE Research Group. (2012b). From writing the grant to working the grant: An exploration of processes and procedures in transition. *Scholarly and Research Communication, 3*(1).

Siemens, L., and INKE Research Group. (2012c). INKE administrative structure: Omnibus document. *Scholarly and Research Communication, 3*(1).

Siemens, L., and INKE Research Group. (2012d). *INKE at the midterm review*. Paper presented at the Research Foundations for Understanding Books and Reading in the Digital Age: E/Merging Reading, Writing, and Research Practices, Havana, Cuba.

Siemens, L., and INKE Research Group. (2012e). Understanding long-term collaboration: Reflections on year 1 and before. *Scholarly and Research Communication, 3*(1), 1-4.

Siemens, L., and INKE Research Group. (2013). Responding to change and transition in INKE's year three. *Scholarly and Research Communication, 4*(3), 12 pp.

Siemens, R. G., Siemens, L., Cunningham, R., Galey, A., Ruecker, S., and Warwick, C. (2012). Implementing new knowledge environments: Year one research foundations. *Scholarly and Research Communication, 3*(1).

Siemens, R. G., Warwick, C., Cunningham, R., Dobson, T., Galey, A., Ruecker, S., Schreibman, S., and INKE Research Group. (2009). Codex ultor: Toward a conceptual and theoretical foundation for new research on books and knowledge environments. *Digital Studies/Le champ numerique, 1*(2).

Spiro, L. (2010). Opening up digital humanities education. Retrieved from http://digitalscholarship.wordpress.com/2010/09/08/opening-up-digital-humanities-education/

SSHRC. (2013). Partnership grants: An overview, December 9, 2013, from http://www.sshrc-crsh.gc.ca/about-au_sujet/partnerships-partenariats/partnership_grants-bourses_partenariats-eng.aspx

The Praxis Program at the Scholars' Lab. (2011a). 2011-12 praxis program charter, September 19, 2012, from http://praxis.scholarslab.org/charter.html

The Praxis Program at the Scholars' Lab. (2011b). 2011-12 praxis program charter Retrieved February 11, 2012, from http://praxis.scholarslab.org/charter.html

The Praxis Program at the Scholars' Lab. (2012). About praxis, October 17, 2012, from http://praxis.scholarslab.org/about.html

The Praxis Program at the Scholars' Lab. (nd). Toward a project charter, September 11, 2013, from http://praxis.scholarslab.org/topics/toward-a-project-charter/

Notes

[1] See R. G. Siemens et al. (2009) for the full grant application.

From Handwritten Text to Structured Data: Alternatives to Editing Large Archival Series

Ronald Sluijter
ronald.sluijter@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

Marielle Scherer
marielle.scherer@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

Sebastiaan Derks
sebastiaan.derks@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

Ida Nijenhuis
ida.nijenhuis@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

Walter Ravenek
walter.ravenek@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

Rik Hoekstra
rik.hoekstra@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

Introduction

One of the key problems in historical political research is that many relevant research questions can only be answered by means of a prolonged and painstaking analysis of large archival series. Questions like: "How did the relationship between the government, the parliament and the political elites change over time?", "What role did political traditions and rituals play?", and "In what ways did the information economy influence the political process?", still require scholars to systematically work through vast bodies of archival material. Only a few scholars, who appeared not to be intimidated by such a daunting task, have come up with long-term analyses of political patterns. This paper proposes a new, digital approach to avoid these time-consuming activities and to open up major archival series for automated text analysis, by applying various tools for text recognition and automated structuring, as well as by using reference data and re-using metadata.

The case study selected to demonstrate the potential of this approach is the opening up of the Resolutions of the Dutch States-General from 1576 to 1795. This archival series of the central assembly of the seven Provinces forming the Dutch Republic is an excellent example of the type of source suitable to answer the above mentioned research questions. Editing the Resolutions has been a task of the Huygens Institute for the History of the Netherlands and its predecessors since 1915 (Japikse and Rijperman, 1915-1970, Van Deursen et al., 1971-1994).[1] This task is hindered, though, by the vast size of the resolutions, which approximately consists of 200,000 pages. The classic edition process, not even aimed at providing a full transcription of the resolutions but only abstracts, reached the year 1625 in 1994. After that, a project was carried out to edit the resolutions from 1626 to 1630 only in digital form, with the help of xml-coding (Nijenhuis, 2006; Nijenhuis et al., 2007). This project ended in 2007 and was not pursued further, because it was clear that this method also was too time and money consuming.

In 2015 we started a totally different approach as an alternative to editing this vast collection of documents. Our goal is to make this important collection accessible, searchable, and analyzable for historical research by applying various advanced digital humanities tools. We will do this by using the metadata of the already printed edited Resolutions, and by enriching and linking the data to other relevant research data. The advantage of this approach is that the Resolutions will be made accessible for researchers in a far quicker way in comparison to the classic way of editing. Also, this approach is to provide insights which will be useful for comparable projects dealing with important archival series in the future, and may provide an alternative to full scale editing of large historical sources in the digital era.

Reusing metadata

On the basis of experience with digital editing, we know that performing a small scale experiment is the best way of establishing best practices and avoiding huge costs. Therefore, we have chosen to apply a multilevel approach with several pilot projects, using various tools which may be applicable for exploring the content of some 200,000 pages of resolutions for historical research. These projects depart from the principle of using what is already there. This means, in practice, that we will construct a framework consisting of the metadata added by the previous editors, like indexes; mark-ups of names, places and institutions; and classification of subjects, as well as contemporary indexes and marginal subject-notes in the resolutions. This framework will serve as a reference to make the resolutions corpus accessible.

Automated Handwritten Text Recognition

Secondly, we experiment with tools for Handwritten Text Recognition (HTR). The software developed by researchers from the Universitat Politècnica de València, which is integrated in the *Transkribus* platform, offers the most valuable results (Sanchez et al., 2013).[2] For the HTR-experiment we manually transcribed 40 pages of handwritten resolutions. Of these pages, 30 were used for training and 10 for testing. The resulting Word Error Rate for this experiment was 40%. We realize that using an only partly correct transcription does not deliver the results one can expect from a traditional edition. The automated transcriptions generated by the HTR-tool should be seen as an alternative that enables researchers to search the text. Of course with a Word Error Rate of 40% this will not deliver perfect results. The HTR may be improved by expanding the training set and by the use of reference data, which is discussed below. In case our approach will be financed for the whole series of resolutions, we intend to use crowdsourcing to improve the HTR results on the handwritten resolutions via the *Transkribus* platform. As has been demonstrated by the developers from Valencia, manual correction of incorrect transcriptions of the HTR-software leads to a recalculation diminishing the mistakes the software made in the rest of the text. This will speed up the work towards an accurate transcription for the whole series. Nevertheless, with the limited number of people able to read 17th-century Dutch handwriting taken into account, we cannot expect crowdsourcing to provide us with a perfect transcription of the whole series of resolutions within a few years.

Automated annotation

Finally, we investigated the use of tools for enriching the resolutions with annotations that will improve exploring the digitized material. For this purpose we used contemporary printed resolutions, of which a series exists from 1703 to 1796. We selected a set of 100 pages from the year 1725, containing 366 resolutions. The text of the resolutions was transcribed and marked up manually using TEI. The printed material consists of blocks of text that can be categorized as follows: "session", consisting of the date of the meeting, the name of the chairman, and the attendees representing each of the seven provinces; "resumption", the summary of the previous meeting; "resolution", the body of the resolutions themselves; and "insertion", mostly incoming letters. We used a standard machine learning approach. The text blocks were marked up with their type manually. Part of the material was used for training the automated categorizer, the other part for testing. The categorizer was trained using as features fixed expressions the successive clerks of the States-General used in their account of the meetings. The comparison of the results with the manual categorization turned out to be promising. We were able to categorize the different types of text blocks with a 98,6% precision.

The next step was to extract information from the text blocks. Building on software of the *Stanford Natural Language Processing Group* we developed a rule-based tool for recognizing dates with a 99,1% precision. The dates in the "session" text blocks could be identified easily, for they have a fixed structure; this allows us to annotate each resolution with its date. Software for identifying more complex dates in the text blocks (for example "the resolution taken yesterday") is yet to be written; it will be used to annotate resolutions with references to other resolutions.

Furthermore, we developed a tool using a Naïve Bayes Classifier for categorizing the attendance list. The manuscript and printed resolutions list the attendants according to the province they represented; the provinces were mentioned in a fixed order. However, at some meetings a province was not represented. With the tool we are able to identify the provinces the attendants as well as the chairman represented in these cases also.

Finally, we took some steps in interpreting the content of the resolutions. Again using fixed phrases and a Naïve Bayes Classifier, it turns out to be possible to recognize in most cases whether or not a decision was taken in a resolution (96,9% precision) and whether the States-General asked a person or a body for advice (99,3%). The results for analyzing to whom the States-General asked advice; and whether and to whom they asked for a report, are yet inconclusive because of the limited amount of training material.

Apart from improving the results for this last mentioned analysis, our future work will be dedicated to several issues. Firstly, applying Named Entity Recognition to the resolutions in combination with the use of reference data. The Huygens ING owns and hosts several relevant data collections, for instance the *Biography Portal of the Netherlands*[3] and the *Compendium of Office-Holders and Civil Servants 1428-1861*[4]with which we will be able to identify persons and institutions mentioned in the resolutions. Secondly, we aim at improving the OCR for the printed resolutions, for the benefit of the automated annotation. Thirdly, we will investigate whether the tools for automated structuring can be applied also to the automated transcription of the handwritten resolutions.

Bibliography

Deursen, A. T. van, Smit, J. G. and Roelevink, J. (Eds.) (1971-1994). *Resolutiën der Staten-Generaal: Nieuwe reeks, 1610-1670*. Vol 7, Gravenhage.

Nijenhuis, I. (2006). Besluiten ontsloten. Resolutiën Staten-Generaal digitaal (1626-1630). *De Zeventiende Eeuw*, **22**: 272-82.

Nijenhuis, I. J. A., De Cauwer, P. L. R., Gijsbers, W. M., Hell, M., Meij, C. O. van der and Schooneveld-Oosterling, J. E. (Eds.) (2007, update 2011). *Resolutiën Staten-Generaal 1626-1630*. Den Haag.

Japikse N. and Rijperman, H. H. P. (1915-1970). *Resolutiën der Staten-Generaal van 1576 tot 1609*. Vol. **14**. Den Haag.

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., Vidal, E. and Does, J. de (2013). tranScriptorium: a European Project on Handwritten Text Recognition. *Proceedings of the 2013 ACM symposium on Document engineering*. New York: ACM Press, pp. 227-28.

Thomassen, T. (2015). *Instrumenten van de macht. De Staten-Generaal en hun archieven 1576-1796*. Den Haag: Huygens ING.

Wittek, P. and Ravenek, W. (2011). Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling. In Maegaard, B. (Ed.), *Supporting Digital Humanities 2011: Answering the unaskable*. Copenhagen.

Notes

[1] http://resources.huygens.knaw.nl/besluitenstatengeneraal1576-1630/index_html_en (accessed 3 March 2016)

[2] https://transkribus.eu/Transkribus/ (accessed 3 March 2016)

[3] http://www.biografischportaal.nl/en/ (accessed 3 March 2016)

[4] http://resources.huygens.knaw.nl/repertoriumambtsdragersambtenaren1428-1861/index_html_en (accessed 3 March 2016)

SMTP: Stedelijk Museum Text Mining Project

Jeroen Smeets
smeetsjeroen@hotmail.com
Maastricht University, Netherlands, The

Johannes C. Scholtes
j.scholtes@maastrichtuniversity.nl
Maastricht University, Netherlands, The

Claartje Rasterhoff
C.Rasterhoff@uva.nl
CREATE, University of Amsterdam, Netherlands, The

Margriet Schavemaker
M.Schavemaker@stedelijk.nl
Stedelijk Museum Amsterdam, Netherlands, The

Introduction

This paper addresses how text-mining, machine-learning and information retrieval algorithms from the field of artificial intelligence can be used to analyze Art-Research archives and conduct (art-) historical research. To gain quick insight into the archive, two aspects are focused on: relations between groups of people using community detection, and global content changes over time using topic modeling. For such archives pre-tagged ground-truth collections are generally not available, and the archives are often too large, geographically distributed, and not always available in digital formats to build such a ground-truth at reasonable costs. To develop and test the validity and relevance of existing tools, close collaboration was established between the AI researchers, museum staff, and researchers in CREATE, a digital humanities project that investigates the development of cultural industries in Amsterdam over the course of the last five centuries.

Data

The research draws on two datasets. The principal dataset is the digitized archive of the Stedelijk Museum Amsterdam, a renowned international museum dedicated to modern and contemporary art and design. The archive of the Stedelijk Museum Amsterdam contains documents from the period 1930-1980. The corpus is a static collection of approximately 160.000 text documents that were digitized using OCR. The second dataset is drawn from Delpher, developed by (Koninklijke Bibliotheek Nederland, 2015). Delpher provides a collection of digitized newspapers, books and magazines that is available for research. A selection of newspapers was made that is used as an additional dataset for this project. Only articles from 1930-1980 that resulted from the query ”Stedelijk Museum” AND ”Amsterdam” were used, forming a set of 18.290 articles.

Methodology

The following methodology uses two approaches to obtain a quick and detailed overview of the content of a digitized archive that contains unstructured information. The first one focuses on the relations between named entities and aims at finding communities in the relation network. The second approach uses time based topic-modeling to get an overview of content changes over time. Finally, a name extraction method is presented that is able to handle multiple causes of name variations.

Relation networks and community detection

In its most basic form, a relation between two named entities can be said to exist when they occur together in the same document. The strength of a relation can be characterized by the number of documents in which both named entities occur. When all the co-occurrences are found, a relation network can be constructed.

In addition, sentiment analysis can be done to further characterize a relation. A sentiment score is assigned to each document, indicating the sentiment content of the document. No distinction is made between positive and negative sentiment polarity. The hypothesis is that relations between individuals with a high sentiment are more interesting than relations with a low sentiment. This is because sentiments around trigger-events are often higher than around common-day events. A lexicon based approach is used with lists of language specific sentiment words. The sentiment score of a document is then given by the sigmoid of the count of the sentiment words in the document, normalized by the number of words in the document.

Finally, community detection algorithms can be applied to the relation network. These types of algorithms aim at finding clusters of groups of entities that have dense connections between members of the clusters and sparse connections with members of other clusters (Fortunato, 2010). The relation weight measure that is used to calculate the communities, is taken as the product of the strength of the relation, i.e. the number of documents where both entities occur in, and the average sentiment score of the documents of a relation. It was found that combining these two measures, resulted in more meaningful communities.

Time based Topic Modeling

In the next approach, topic modeling algorithms are applied to analyze the information content and their evolution over time. Topic modeling tries to discover the underlying thematic structure in a collection of documents. Non-Negative Matrix Factorization (NMF) is being used as a tool for topic modeling (Arora et al., 2012). NMF is an unsupervised method where a matrix is approximated by two low rank non-negative matrices. The extracted semantic feature vectors have only non-negative values and are sparse so they are easily interpretable. Furthermore, NMF is shown to generate more consistent results over multiple runs (Choo et al., 2013), compared to other tools used for topic modeling such as LDA (Blei et al., 2003).

The approach suggested in (Vaca et al., 2014) uses a time-based collective matrix factorization based on NMF and is used in this project. It extends NMF by introducing a topic transition matrix that allows to track topics as they emerge, evolve and fade over time.

Name Extraction

The following method was used to extract named entities from a collection of documents in order to build the relation network. It handles different causes of name variations such as OCR induced errors commonly found in digitized document collections, spelling mistakes, name abbreviations and first and last name combinations.

The method makes use of lists of name variations. Starting from a set of names extracted from a name database, such as RKDArtists and (RKD, 2015), the document collection is searched for possible name variations. These variations are found by searching for the last name using a fuzzy search. The similarity between the group of tokens around the found last name, and the original name is then calculated as a similarity score. The similarity score calculation is based on the idea described in (Song and Chen, 2007), which uses a n-gram set matching technique. The lists of name variations can then be evaluated manually or a threshold on the similarity score can be used to identify name variations that correspond to the original name. The method using a threshold of 0.9 on the similarity score was tested on 50 randomly chosen names. The average precision was found to be 81 percent.

Results

A relation network was constructed for the document collection of the archive of the Stedelijk Museum Amsterdam. Only artists with the graphic artist qualification in the RKDArtists and database were used. The methods were implemented using available open source software libraries such as the Apache Lucene text search engine library (The Apache Software Foundation, 2015) and the Gephi platform (Bastian et al., 2009). The standard community detection feature in Gephi was used, which is based on the Louvain method (Blondel et al., 2008). The result is shown in Figure 1. The color of the relation between the nodes indicates the average sentiment score of the relation, starting from blue (neutral) to red (high sentiment content). Communities such as group exhibitions, art movements or a group of artists closely related to the museum director, could be identified with the help of a museum expert.

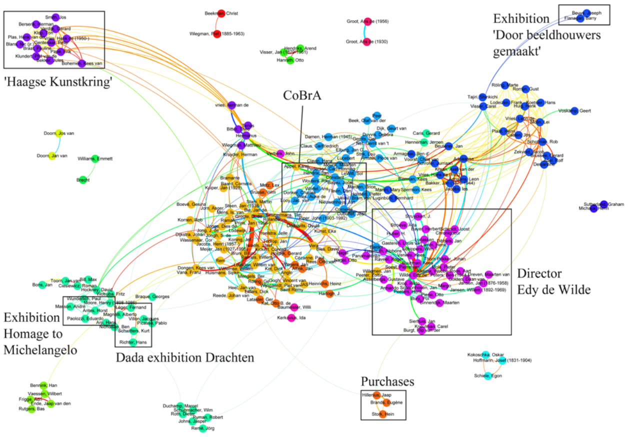


Figure 1: Found communities for graphic artists in the archive of the Stedelijk Museum

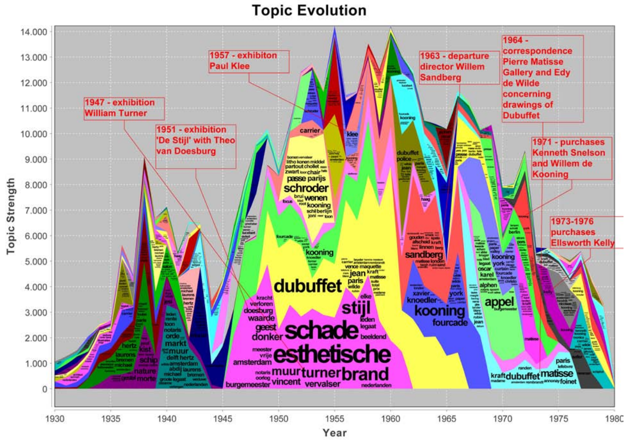


Figure 2: Time based topic modeling for the archive of the Stedelijk Museum Amsterdam

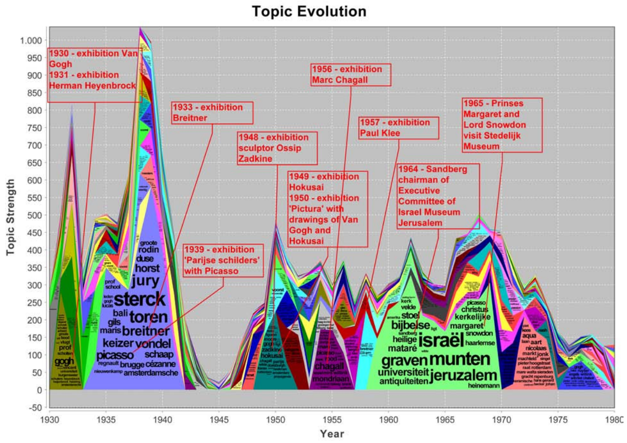


Figure 3: Time based topic modeling for Delpher newspaper articles

The time based topic modeling algorithm suggested in (Vaca et al., 2014) was implemented in MATLAB and Java. The algorithm was applied to both the archive of the Stedelijk Museum Amsterdam and newspaper articles from the Delpher database. The results are visualized over time in the form of stacked topic rivers (Wei et al., 2010), shown in Figure 2 and Figure 3. Several exhibitions and events could be identified and are annotated on the chart.

Conclusion

This paper discusses two approaches to gain insight into a digitized archive. Relation networks of persons with community detection are considered, relying on a robust name extraction method. Furthermore, the evolution of content over time can be explored using time based topic modeling.

For the humanities researchers in this project, the main aim was to asses the research potential of computational analysis of digitized art archives in general, and the Stedelijk Museum in particular. Two types of preliminary research questions were developed to do so. The first type had to do with identifying patterns of change and continuity, across time and place. These include for instance tracing the position of the Stedelijk Museum as an intermediary in Dutch design industries, or the development of the Stedelijk Museum as an increasingly international player. The second type of question is less concerned with general historical patterns, and more with specific art-historical research questions, regarding for instance (networks of) particular artists, artworks or exhibitions. But before we could start asking such questions to digitized art-historical archives, the quality and accessibility of the texts needed to be established. Secondly, specific methods needed to be explored and adapted in order to clean, identify, retrieve, extract, and structure the texts. The first results presented in this paper demonstrate that even though they may not be clean at the first try or capture all historical nuance, they do help archives to open up and show unexpected relationships and patterns, to answer specific questions, and to get connected with other relevant sources, such RKDartists and Delpher. The community detection in relation with sentiment mining, the topic modeling and name extraction method developed in this project therefore provide a solid basis for the next step in assessing the research potential of art-historical archives: developing in-depth case studies, again in close collaboration with art-historians and historians, allowing the archive to speak up in unprecedented ways, offering access to hidden story lines that subvert and augment prevailing historical narratives.

Bibliography

Arora, S., Ge, R. and Moitra, A. (2012). Learning topic models - going beyond SVD. *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*. IEEE, pp. 1–10.

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*, **8**: 361–62.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**: 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10): P10008.

Choo, J., Lee, C., Reddy, C. K. and Park, H. (2013). Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *Visualization and Computer Graphics, IEEE Transactions on*, **19**(12): 1992–2001.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**(3): 75–174.

Koninklijke Bibliotheek Nederland (2015). Delpher - Boeken Kranten Tijdschriften http://www.delpher.nl/ (accessed 1 November 2015).

RKD (2015). Netherlands Institute for Art History https://rkd.nl/en/ (accessed 1 November 2015).

Song, S. and Chen, L. (2007). Similarity joins of text with incomplete information formats. *Advances in Databases: Concepts, Systems and Applications*. Springer, pp. 313–24.

The Apache Software Foundation (2015). Apache Lucene - Welcome to Apache Lucene http://lucene.apache.org/ (accessed 1 November 2015).

Vaca, C. K., Mantrach, A., Jaimes, A. and Saerens, M. (2014). A time-based collective factorization for topic discovery and monitoring in news. *Proceedings of the 23rd International Conference on World Wide Web*. ACM, pp. 527–38.

Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., Shi, L., Tan, L. and Zhang, Q. (2010). Tiara: a visual exploratory text analytic system. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 153–62.

Comparing Digital Scholarly Editions

David Neel Smith
nsmith@holycross.edu
College of the Holy Cross, United States of America

Sephanie Lindeborg
smlind13@g.holycross.edu
College of the Holy Cross, United States of America

Citing and comparing texts

Explicit and tacit assumptions of traditional text criticism have been questioned for decades,[1] but the creation of digital scholarly editions has provoked discussion ranging from practical questions of method, to theoretical debate about what constitutes the nature of an edition in an electronic environment.[2] In this paper, we address questions of what it means to compare two scholarly editions, and demonstrate applications of our approach to compare manuscripts of the *Iliad*.

One characteristic of a scholarly edition in any medium is that it makes a text canonically citable. Canonical citation is an essential prerequisite: in order to compare more meaningful units than streams of characters, we must be able to identify and align passages in different versions of a text. We use the technology-independent CTS URN notation[3] to identify books and lines of the *Iliad*. CTS URNs could be applied to any citation scheme, but logical schemes (such as those typically used to cite biblical passage by chapter and verse, or legal document by numbered section and subsection) are superior to arbitrary schemes based on physical artifacts (such as citing Plato by Stephanus page, or referring to the physical page of a specific edition of the works of Jane Austen) since they ensure that our comparison is organized in chunks of texts meaningful for scholarly analysis.

A model for text comparison

Traditional Homeric scholarship offers a useful model for comparing aligned citable texts. Homerists use the terms "vertical difference" and "horizontal difference" to describe two kinds of variety: "vertical difference" refers to entire lines that are present in one text but absent in the other, or that occur in a different sequence in the two texts. "Horizontal difference" refers to lexical differences within a single line. We can generalize the two dimensions of this approach, and understand our comparison of texts as determining the structural and lexical variation between texts.

Structural differences are simply differences in citation structure. For texts cited by CTS URNs, then, we can reduce the determination of structural differences to a comparison of ordered lists of each document's CTS URNs.

Lexical differences are differences in the readings within a single citation unit (line of the *Iliad*, subsection of a legal text, etc.) The crucial question is: what produces a "reading"? Simply comparing streams of characters, or assuming that a stream of characters can be parsed into tokens based on some criterion such as splitting word tokens by white space or punctuation is dangerously underconceptualized. Instead, we recognize that any analysis that tokenizes a citable unit of text for a specified purpose produces an ordered list of tokens that we can compare in order to determine lexical differences between two texts. The lexical type of the token will depend on the goal of the comparison. As we subsequently illustrate, for example, we could analyze literal textual tokens, orthographically normalized tokens, or even morphologically or metrically analyzed tokens.

Implementing the model

When we determine structural ("vertical") variation by comparing ordered lists of URNs and lexical ("horizontal") variation by comparing ordered lists of tokens for each citable unit of text, we are performing exactly the same operation: comparison of ordered lists. This is one of the most studied and best understood problems in computer science, and a typical exercise in first-year programming courses. We have implemented the standard algorithm for Longest Common Subsequence (or LCS)[4] in a library freely available in source code or binary .jar for JVM languages. In addition to solving the LCS, the library determines what items appear in one list but not the other, and what items appear in both lists but in a different order.

Applying the model to the *Iliad*

We illustrate the possibilities of this approach by repeatedly comparing incompletely published manuscripts of the Iliad, focusing especially on *Iliad* 8 in two manuscripts, in the Biblioteca Marciana in Venice and in the Escorial monastery near Madrid. All of our comparisons find the same structural differences. (The run of lines from *Iliad* 8.466-8.468 is present in some manuscripts, for example, but absent from others.) The lexical comparisons, on the other hand, vary depending on the features we analyze.

We begin with a simple tokenization of the literal diplomatic text split on white space. Inventorying the tokens in each manuscript is essentially the collation phase of a traditional edition, but when we fully account for differences in punctuation, accent, abbreviation and spelling, the vast quantities of differences between manuscripts informs us about aspects of Byzantine orthographic practice that are normally suppressed in critical editions.

We next tokenize the same text to a normalized orthography eliminating punctuation, and adapting both accents and spelling to modern conventions. This comparison most closely approaches what we find in a typical critical edition (except that its listings of tokens present, absent or relocated in different manuscripts is comprehensive, rather than selective). Viewed from this perspective with orthographic differences removed, we find much greater agreement in the text of the Venice and Escorial manuscripts, although we still find passages like 8.137 where the reins of Nestor's chariot are either "shining" (σιγαλόεντα) or "red-purple" (φοινικόεντα).

This comparison also reports differences in passages like *Iliad* 9.3, where a few manuscripts have βεβλήατο against the majority with βεβολήατο. The "differences" are actually equivalent forms of the same verb (an epic pluperfect of βάλλω). Depending on our interests, we might prefer to view these two literal variants as identical. We next tokenize the text not to representations of the specific form found in the text, but instead to the lexical entity ("dictionary form") from which the word derives. In this tokenization, the same lexical entity is given for each of the two variant forms, and the lines are, by this reading, equivalent.

Since the formulaic variation illustrated by different readings for the same passage is metrically conditioned, we might also want to tokenize the text to metrical units. Like the preceding tokenizing to an abstract lexical entity, this is often considered beyond the scope of a critical edition, but we do not need to make any procedural distinction in our digital comparison. Reading the same line 9.3 metrically, for example, we next tokenize the dactylic hexameter into six metrical feet. In the majority manuscripts with βεβολήατο, we "read" the text with a dactyl in the third foot,

—∪ ∪ —— — ∪ ∪ — ∪ ∪ — ∪ ∪ —×

while in the minority manuscripts with βεβλήατο we read a spondee

— ∪ ∪ —— —— — ∪ ∪ — ∪ ∪ —×

Each of these comparisons captures a distinct feature of the text. In every case, the analyses are keyed to the CTS URN of the text they analyze, so we can readily combine and compare the results of distinct analyses. In the Iliadic examples, we could equally easily identify passages that are metrically identical with either different vocabulary items or different forms of the same vocabulary item; or, as in 9.3, metrically distinct passages with identical vocabulary in different forms.

Rethinking digital editions

We recognize, as others have before, that many assumptions in the traditional practice of critical editing are self-contradictory and unnecessary in a digital environment. Digital editors are not constrained to eliminate the evidence of manuscripts judged not valuable (*eliminatio codicum descriptorum*); they do not have to select only significant variants (*selectio*) based on the evaluation of a limited set of crucial passages (*examinatio locorum criticorum*); they do not have to present material supporting their editorial choices in a critical apparatus that is flawed both by its circular logic of selectively publishing evidence and by its notational deficiency (a deficiency that has been clearly recognized only when the apparatus is computationally processed).[5] One of the most significant consequences of working with scholarly editions in a digital environment is the potential to automate systematic and comprehensive comparisons of various classes of features across a set of full diplomatic editions.

The comparisons of Iliadic manuscripts presented here further show that the analysis underlying a traditional critical edition is functionally no different than any other kind of analytical comparison: critical editing is one approach to analyzing a comparable set of texts. Our model of textual comparison compels us to specify unambiguously the process that generates our sequence of lexical tokens. This permits us to apply completely generic tools for comparing ordered lists, and to construct increasingly complex cascades of aligned analyses.

Bibliography

Boschetti, F. (2007). Methods to extend Greek and Latin corpora with variants and conjectures: Mapping critical apparatuses onto reference text. In *Proceedings of the Corpus Linguistics Conference.CL2007.* Birmingham. http://ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf (accessed March 4, 2016).

Pasquali, G. (1934). *Storia della tradizione e critica del test.* Florence.

Pierazzo, E. (Ed.). (2015). *Digital Scholarly Editing: Theories, Models and Method.* Ashgate.

Sutherland, K. and Deegan, M. (Eds.). (2009). *Text Editing, Print and the Digital Worl.* Ashgate.

Notes

[1] As early as 1934, Pasquali already pointed out that the necessary assumption of the "Lachmannian method" that each copy derives from a single archetype is demonstrably wrong in many instances (Pasquali, 1934).

[2] See for example the collection of essays edited by Sutherland and Deegan (Sutherland and Deegan, 2009), or the recent broad survey assembled by Pierazzo (Pierazzo, 2015).

[3] On CTS URNs, see http://cite-architecture.github.io/ctsurn/.

[4] https://en.wikipedia.org/wiki/Longest_common_subsequence_problem

[5] An important but largely unrecognized implication of Federico Boschetti's work parsing a critical apparatus of Aeschylus is that more than 10% of the entries in the critical apparatuis were not clearly enough expressed to be correctly mapped on to the section of the main text they annotate. This was not due to lack of diligence by the editors: rather, it reflects the notational ambiguity of the traditional apparatus. (Boschetti, 2007)

Evaluating GitHub as a Platform of Knowledge for the Humanities

Lisa Spiro
lisamspiro@gmail.com
Rice University, United States of America

Among the most popular platforms for digital humanities development projects is GitHub.[1] GitHub provides web-based hosting for coding and other collaborative projects, building on the Git version control system. Founded in 2008, GitHub now hosts approximately 31 million repositories and 12 million users, making it " the largest online storage space of collaborative works that exists in the world" (GitHub, 2016; Orsini, 2013). Using GitHub, developers can fork (copy) a public repository to their own account, change the code, and submit a pull request to share the modifications with the repository owner, who can then "merge" the code with the original. GitHub also enables users to create profiles, "follow" others, "star" projects and "watch" them evolve, serving as a social network for developers (Brown, 2014).

Digital humanities (DH) researchers are drawn by GitHub's support for version control and collaboration, as well as by its free hosting for publicly available projects. A range of digital humanities projects use GitHub, including: Software

- Writing projects
- Taxonomies and community documentation
- Websites
- Datasets
- Course materials
- Syllabi
- Research notes

What digital humanities researchers are adopting GitHub, and why? What are the benefits and risks of employing GitHub for digital humanities work?

Academia's growing reliance on GitHub requires careful consideration. As a for-profit company, GitHub does not necessarily operate with the interests of academics at heart. Yet it provides services that would be difficult for scholars to secure themselves, and it enables collaboration. To develop an informed view of GitHub and similar services, we need to establish clear criteria for evaluating platforms. In choosing a platform for a web project, Quinn Dombrowski recommends considering functionality, familiarity, community, support and cost (Dombrowski, 2013). We would add support for openness and sustainability as core criteria for digital humanities platforms.

Through an initial case study of GitHub, we will examine these criteria for evaluating DH platforms:

- **Functionality:** GitHub offers several features that make it attractive to researchers, particularly open science advocates. Karthik Ram suggests that Git (and by extension GitHub) supports open science by providing decentralized version control; attributing changes to authors; supporting distributed backup of data; enabling projects to branch in new directions; collecting feedback through issue trackers; and facilitating reuse through forking (Ram, 2013). Likewise, Konrad Lawson touts the power of GitHub in facilitating "collaboration without collaboration" (easily modifying someone else's code through forking, and contributing that code back through a pull request) and detailed credit for contributions (Lawson, 2013c; Lawson, 2013a). While GitHub can be used for a range of texts, from syllabi to code, it's not necessarily well suited for all uses. For example, Mark Sample notes the significant labor and potentially low rewards in putting syllabi into GitHub (Sample, 2012a). Lawson observes that using GitHub for writing projects requires overcoming a fairly steep learning curve, using plain text (or a converter), creating short documents, and dealing with limited support for non-textual files (Lawson, 2013d).
- **Familiarity/ease of use:** As Lincoln Mullen points out, GitHub's learning curve poses a barrier to entry for some potential collaborators (Mullen, 2012). Indeed, participants discussing how to make it easier for women to contribute to *Programming Historian* identified the publication's reliance on GitHub as an obstacle (Crymble et al, 2015). To what extent do the challenges of using GitHub limit its adoption in the humanities?
- **Community:** As more digital humanists adopt GitHub, it becomes even more attractive, since you are more likely to find collaborators and to gain recognition for your work. Yet the wide adoption of GitHub may reduce diversity and increase dependency on a commercial platform. Not everyone wants to participate in this community. Lawson points to several social and cultural obstacles to GitHub enabling richer academic collaboration, including reluctance to embrace "forking" as means of building on another's work; fears of plagiarism; concerns that the original voice of the author will be lost; anxiety that transparency will reveal one's scholarly flaws; and worry that ideas will be stolen or misused (Lawson, 2013b).
- **Support:** With such a large community, new users can turn to a number of resources to learn how to use GitHub. Some university IT groups offer limited support for GitHub, but in general users are left to secure their own support.
- **Cost/business model:** GitHub uses a "Freemium" business model in which it hosts public repositories for free and charges companies for private repositories (Brown, 2014). It also offers up to five free private repositories to academic researchers and twenty to research groups. While free holds appeal, should the digital humanities community be concerned about becoming dependent on a platform developed by a for-profit company? As Sample warns, "History suggests that relying too much on a commercial service with interests that do not neces-

sarily align with our own is no way to sustain the work of the humanities"(Sample, 2012b). GitHub has attracted $350 million in venture capital and is now valued at about $2 billion, so it faces pressure to generate a profit (Gage, 2015). Klint Finley argues that GitHub's business interests may work against its open source mission, pointing to SourceForge as an example of an open source software site that went astray (Finley, 2015). When SourceForge was acquired, it began to display junky third-party ads that misled people into downloading malicious software, prompting projects such as GIMP and VLC to leave. While GitHub is not funded through ad revenue, its business model could change under pressure from investors.

- **Support for openness:** Compared to some web platforms that claim user-produced content as their own, GitHub articulates an open approach to intellectual property: "Your profile and materials uploaded remain yours"(GitHub, 2015). But should we be concerned about clauses reserving the right to remove content and requiring users to defend and indemnify GitHub against suits alleging that their content violates the law? Does GitHub's model of providing free public repositories lead some users to share work that they otherwise would keep private?
- **Sustainability:** GitHub is not meant to be a preservation repository, and it is easy to delete a public repository (Bergman, 2012). However, Git's distributed, decentralized approach to versioning provides protection against data loss, since everyone who contributes to a GitHub project has a local copy of the code (Finley, 2015).

In addition to these criteria, we will also consider the significance of factors such as accessibility and multilingualism.

In performing this research, we are first identifying digital humanities users by 1) searching for publicly available GitHub accounts associated with presenters at the last three Digital Humanities conferences and 2) searching for GitHub accounts associated with Digital Humanities centers listed on CenterNet. To understand patterns of collaboration and code reuse, we will analyze publicly available statistics for selected users such as number of commits, branches, releases and contributors, as well as networks connecting users. We will survey GitHub DH users to understand how and why they use GitHub, its strengths, and its weaknesses. We will also conduct interviews with selected GitHub users. Where possible, we will use GitHub to share ongoing work about this project.[2] By analyzing public GitHub statistics and gathering insights and information from users, we will illustrate how GitHub is being used in the digital humanities community and develop principles for evaluating platforms.

Bibliography

Bergman, C., (2012). On the Preservation of Published Bioinformatics Code on Github. *An Assembly of Fragments*. Available at: https://caseybergman.wordpress.com/2012/11/08/on-the-preservation-of-published-bioinformatics-code-on-github/ (accessed 1 November 2015).

Brown, M. (2014). GitHub - Cracking the Code to GitHub's Growth. *GrowthHackers*. Available at: https://growthhackers.com/growth-studies/github (accessed 13 October 2015).

Crymble, A., Posner, M., et al. (2015). How Can We Make The PH More Friendly For Women To Contribute? Issue #152. *Programming Historian*. Available at: https://github.com/programminghistorian/jekyll/issues/152 (accessed 12 February 2016).

Dombrowski, Q. (2013). Choosing a platform for your project website. *Berkeley Digital Humanities*. Available at: http://digitalhumanities.berkeley.edu/blog/13/12/04/choosing-platform-your-project-website (accessed 14 October 2015).

Finley, K. (2015). The Problem With Putting All the World's Code in GitHub. *WIRED*. Available at: http://www.wired.com/2015/06/problem-putting-worlds-code-github/ (accessed 13 October 2015).

Gage, D. (2015). GitHub Raises $250 Million at $2 Billion Valuation; Capital Raise Puts Company's Total Funding at $350 Million. *Wall Street Journal* (Online), 29 July.

GitHub. (2015). Github Terms Of Service - User Documentation. *Help.github.com*. Available from: https://help.github.com/articles/github-terms-of-service/ (accessed 28 October 2015).

GitHub. (2016). Press. Available at: https://github.com/about/press (accessed 5 March 2016).

Lawson, K. M. (2013a). File and Repository History in GitHub. *The Chronicle of Higher Education Blogs: ProfHacker*. Available at: http://chronicle.com/blogs/profhacker/file-and-repository-history-in-github/48047 (accessed 30 October 2015).

Lawson, K. M. (2013b). Fork the Academy. *The Chronicle of Higher Education Blogs: ProfHacker*. Available at: http://chronicle.com/blogs/profhacker/fork-the-academy/48935 (accessed 30 October 2015).

Lawson, K.M. (2013c). Getting Started With a GitHub Repository. *The Chronicle of Higher Education Blogs: ProfHacker*. Available at: http://chronicle.com/blogs/profhacker/getting-started-with-a-github-repository/47393 (accessed 30 October 2015).

Lawson, K.M. (2013d). The Limitations of GitHub for Writers. *The Chronicle of Higher Education Blogs: ProfHacker*. Available at: http://chronicle.com/blogs/profhacker/the-limitations-of-github-for-writers/48299 (accessed 30 October 2015).

Mullen, L. (2012). How ready are DHers to use GitHub for non-code projects? *Digital Humanities Questions and Answers*. Available at: http://digitalhumanities.org/answers/topic/how-ready-are-dhers-to-use-github-for-non-code-projects (accessed 30 October 2015).

Orsini, L. (2013). GitHub For Beginners: Don't Get Scared, Get Started. *ReadWrite*. Available at: http://readwrite.com/2013/09/30/understanding-github-a-journey-for-beginners-part-1 (accessed 13 October 2015).

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, **8**(1): 7.

Sample, M. (2012a). Git a Fork in My Syllabus, It's Done. *The Chronicle of Higher Education Blogs: ProfHacker*. Available at: http://chronicle.com/blogs/profhacker/git-a-fork-in-my-syllabus-its-done/40331 (accessed 30 October 2015).

Sample, M. (2012b). GitHub Fever. *Digital Culture Week*, **1**(3). Available at: http://www.digitalculture.org/2012/06/08/dcw-volume-1-issue-3-distant-and-familiar/ (accessed 31 October 2015).

Notes

[1] This work emerges out of Rice University's John E. Sawyer Seminar on Platforms of Knowledge in Wide Web of Worlds, supported by the Andrew W. Mellon Foundation.

[2] https://github.com/lms4w/githubproject

A Study of Knowledge Integration in Digital Humanities Based on Bibliographic Analysis

Muh-Chyun Tang
muhchyun.tang@gmail.com
Dept. of Librarya and Information Science, National Taiwan University, Taiwan, Republic of China

Yun Jen Cheng
yjcheng0314@gmail.com
Dept. of Librarya and Information Science, National Taiwan University, Taiwan, Republic of China

Kuang Hua Chen
khchen@ntu.edu.tw
Dept. of Librarya and Information Science, National Taiwan University, Taiwan, Republic of China

Jieh Hsiang
jieh.hsiang@gmail.com
Dept. of Computer Science and Information Engineering
National Taiwan University

This is an extension of our previous work in which a longitudinal (1989-2014) co-citation analysis of literature in Digital Humanities was (DH) conducted to explore the degree of interdisciplinarity in DH. The selection of the literature was based on a combination of keyword search with Scopus and articles published in the key journals in the field (i.e. journals published by members of ADHO). A dual trend of gradual increases in both topical diversity and network cohesion, two hallmarks of interdisciplinarity (Porter et.al., 2007; Rafols and Meyer 2010), were found in the co-citation network. For example, as shown in Figure 1, the average path length remained steady relative to the gradual increase of the network diameter, which suggests continuing growth of the literature and its gradual consolidation (the sharp rise at the tail being the artifact of citation window). On the other hand, the growing diversity was demonstrated by the steady increase of distinct author-assigned keywords over time.

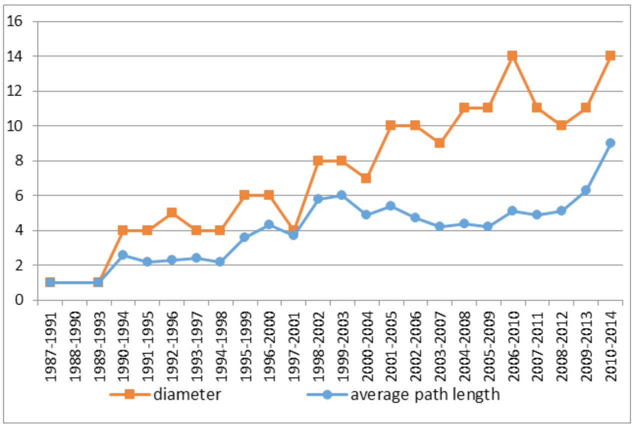


Figure 1. Growing of co-citation network and average path length

In this article, the results of further analyses were reported that aimed to, firstly, further explore of the issue of disciplinary cohesion using co-authorship and bibliographic coupling networks; and secondly, to identify cohesive subgroups or specialties in the DH literature.

Disciplinary cohesion of DH

Based on the assumption that the knowledge integration process in research communities depends heavily on the topology of the underlying social network, Moody used the structure of collaborative (i.e. co-authorship) network to represent the integration of knowledge in Sociology over time (Moody, 2004). The concept of "structure cohesion" or "connectivity" in network analysis was used to measure the degree of social cohesion in Sociology, which is defined as "the extent to which a network will remain connected when nodes are removed from the network (Moody and White, 2003)." Moody (2004) discussed three types of network structures: star production, small world, and structurally cohesive and surmised on the corresponding collaborative practices each represents. Moody (2004) believe that a cohesive collaboration network signals the presence of "permeable theoretical boundaries and generic methods" that allows scholars specialized in particular theoretical, empirical or theoretical skills to collaborate freely. He added that, if enough scholars engage in this kind of cross-fertilization, mixing across multiple areas, there will be few clear divisions presented in the collaborative network (Moody, 2004). Similarly, Carolan (2008) used network structure of articles published by a leading journal in Education to examine how well the heterogeneous set of ideas and practices were integrated within the discipline.

As shown in Figure 2, contrary to the co-citation network, the DH co-authorship network is very sparse and highly fragmented. The percentage of the nodes in the main component hovered only below 20% even after discounting the isolates, which is extremely low compared to other disciplines or research areas (See Table 1). Notice the contrast is especially striking with sciences, medicine, and IT. The low percentage of nodes in main component (Figure 2), coupled with extremely high clustering coefficient and modularity (Figure 3), indicated most collaboration took place at the local level, lacking global "shortcuts" found in the small-world model to hold the network together (See Figure 4).

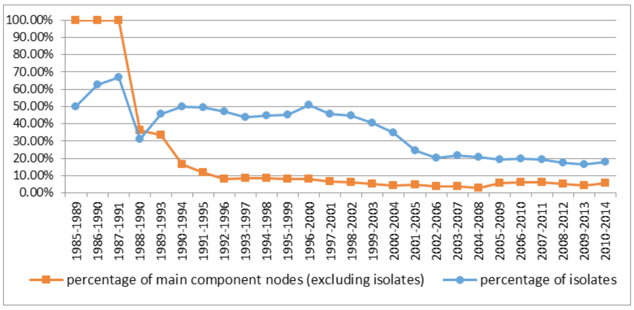


Figure 2. Percentage of nodes in the main component in the co-authorship network

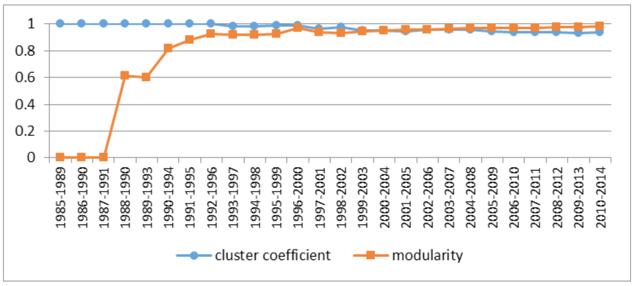


Figure 3. Trends of clustering coefficient and modularity in co-authorship

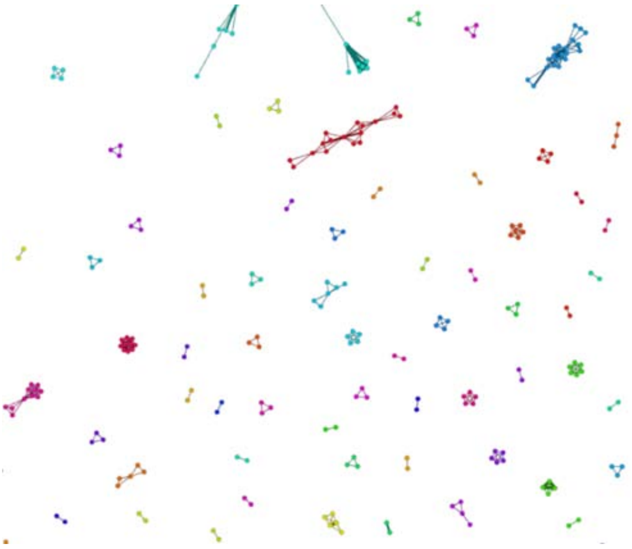


Figure 4. Part of the co-authorship network

A closer examination of the co-authorship network shows that, beyond the two largest components, there were relatively few international collaborations (See Figure 5). The largest component was composed of scholars mainly from the U.S. (28.53), Canada (27.12%), U.K. (26.84%), and Germany (10.45%); the second largest component was composed of scholars from the U.S. (39.71%), the Netherlands (30.88%), and Japan (5.88%); and the third component was composed of all Italian scholars.

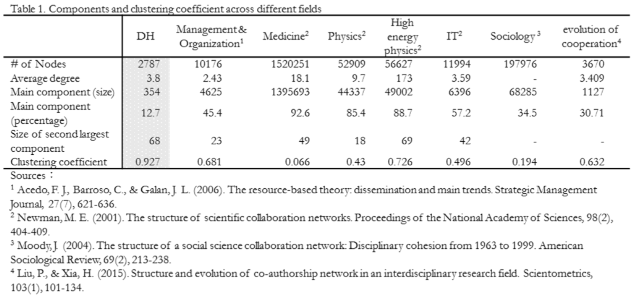


Table 1. Components and clustering coefficient across different fields

	DH	Management & Organization[1]	Medicine[2]	Physics[2]	High energy physics[2]	IT[2]	Sociology[3]	evolution of cooperation[4]
# of Nodes	2787	10176	1520251	52909	56627	11994	197976	3670
Average degree	3.8	2.43	18.1	9.7	173	3.59	-	3.409
Main component (size)	354	4625	1395693	44337	49002	6396	68285	1127
Main component (percentage)	12.7	45.4	92.6	85.4	88.7	57.2	34.5	30.71
Size of second largest component	68	23	49	18	69	42	-	-
Clustering coefficient	0.927	0.681	0.066	0.43	0.726	0.496	0.194	0.632

Sources：

[1] Acedo, F. J., Barroso, C., & Galan, J. L. (2006). The resource-based theory: dissemination and main trends. Strategic Management Journal, 27(7), 621-636.

[2] Newman, M. E. (2001). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98(2), 404-409.

[3] Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. American Sociological Review, 69(2), 213-238.

[4] Liu, P., & Xia, H. (2015). Structure and evolution of co-authorship network in an interdisciplinary research field. Scientometrics, 103(1), 101-134.

Figure 5. Three largest components in the co-authorship network

The identification of specialties in DH

Modularity maximization graph partition was applied to both the co-citation and bibliographic coupling networks to identify subgroups or specialties in DH. Both co-citation and bibliographic coupling have been widely used to establish similarity or linkages between documents in bibliometrics (See, for example, Yan and Ding, 2012; Boyack and Klavans, 2010).

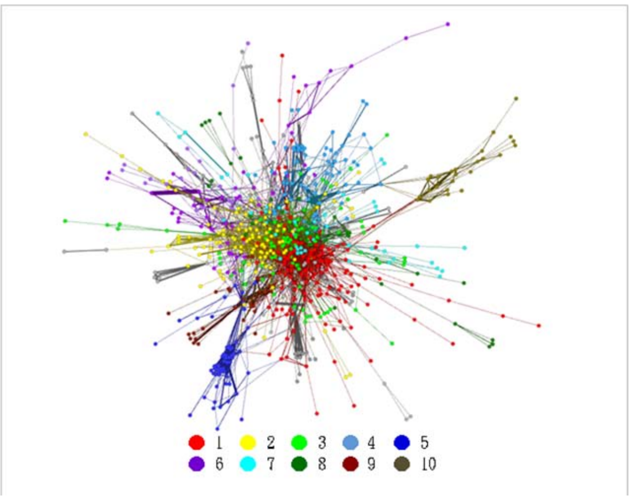


Figure 6. Modularity analysis of co-citation network

The co-citation network was construed by using Google Scholar's citation tracing function. The citations received by every article in our target set were first identified and downloaded then pair-wise matching was performed to identify shared citations. The bibliographic coupling was generated by pair-wise comparison of cited references retrieved from Scopus. A threshold of shared 3 citations in the reference lists was set to dichotomize the network. Figure 5 and 6 show the results of the modularity-based partition resulted from co-citation and bibliographic coupling networks, respectively. Due to the lack of global cohesion, only the giant component in either network was analyzed.

We are currently in the process of identifying the research topics represented by the clusters in either network, which will be done by examining the author-assigned keywords and authors appearing in each cluster. The labeling of the clusters by each's prominent authors has been shown to be an effective way of visualizing a field (e.g. White and McCain, 1998). Interviews with experts who have broad knowledge in the field of DH will be done in order to help us interpret the meaning of the clusters. Efforts will also be made to explore the reasons behind the lack of cohesion in the DH co-authorship network.

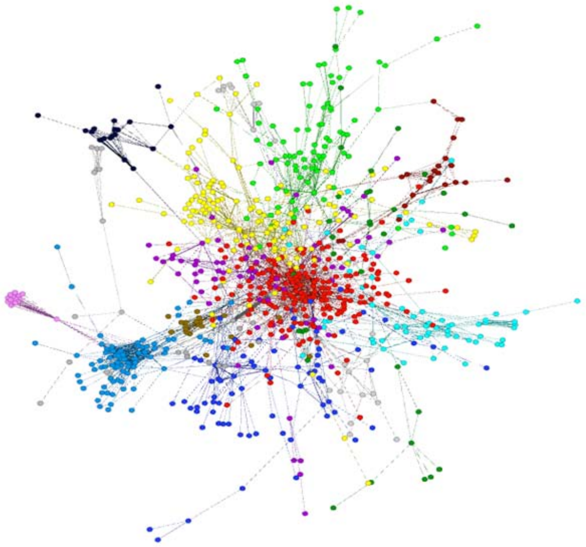


Figure 7. Modularity analysis of bibliographic coupling network

Bibliography

Boyack, K.W., Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, **61**(12): 2389-404.

Carolan, B.V. (2008). The structure of educational research: The Role of Multivocality in Promoting Cohesion in an Article Interlock Network. *Social Networks*, **30**(1): 69-82.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American sociological review*, **69**(2): 213-38.

Moody, J. and White, D.R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, pp. 103-27.

Porter, A.L., Cohen, A.S., Roessner, J.D. and Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, **72**(1): 117-47.

Rafols, I. and Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience, **82**(2): 263-87.

White, H.D. and McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American society for information science*, **49**(4): 327-55.

Yan, E. and Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, **63**(7): 1313-26.

Computers on Law and Order

Jeff Thompson
mail@jeffreythompson.org
Stevens Institute of Technology, United States of America

Detailed accounts have been written of mainframes and cloud computing, social media and online commerce, but there are few books about the more humble aspects of technological culture. Screensavers, bubble-jet printers, computer desks, and other physical technologies are thrown in the trash or overwritten with new versions; the way we talk about computers, the "Web," and the ways technology shapes culture has changed considerably since the birth of the PC. This paper examines how we can find anthropological details about our relationship with technology through popular media, specifically the television program *Law and Order*.

In 2012, I was commissioned by the new media arts organization Rhizome to create a project recording every computer on *Law and Order*. After watching all 319 hours of the show (or the equivalent of about two straight months watching 40-hours a week) and extracting approximately 11,000 screenshots of computers and related technologies, it is clear that *Law and Order* forms a unique database of images and speech, and one that reflects the fascinations, fears, and biases of its time. *Law and Order's* long run and its "ripped from the headlines" content makes it a useful lens with which to look at a period of great political and economic change in the United States. In particular, the show coincides with a major cultural shift: the rise and eventual ubiquity of computers and networked technologies over a crucial 20-year period in technological history.

Using my *Computers on Law and Order* project as a case study, this paper focuses on how these kind of details that can be unearthed from popular media, and that in fact such sources may be the only way to recover the most mundane details. Through a series of categorized screenshots and quotations, I examine several pathways through the archive of the show: the physical infrastructure of computers from shared desktop terminals to smartphones, the development of software interfaces from often-faked text-only input to interactive graphical user interfaces, and peripherals such as mice and printers.

The project can be viewed at: http://www.computersonlawandorder.tumblr.com

Thresholds: Valuing the Creative Process in Digital Publishing

Whitney Trettien
trettien@email.unc.edu
UNC Chapel Hill

Frances McDonald
fran.mcdonald1@gmail.com
Duke University

Introducing *thresholds*

handwritten sticky notes, highlighted document pages, and grainy photographs rub against one another, forming dense and shifting thickets. the blank spaces between once-distinct districts become cluttered and close. geographically distant realms ache to converge. the bookcase furiously semaphores toward the far corner of the room. thin lines of colored paper arrive to splay across sections. the wall bursts at every seam.

Whether it be real or virtual, every project has its own "wall": the irrepressibly interdisciplinary network that inspires and propels the work. Populating this capharnaum are the ideas, images, sentences, scenes, and characters that "stick to us," to use Lara Farina's evocative phrase (Farina, 2014). They are the "encounters" that Deleuze describes as the impetus toward work, the things that "strike" us, as Benjamin puts it, like a hammer to unknown inner chords (Deleuze, 1988; Benjamin, 1999). This affective principle of collection (what strikes you) means that the wall is an intensely personal artifact. Its unique architecture springs from a thinker's nomadic wanderings through and amidst a cultural and aesthetic landscape, whose dimensions are stretched beyond traditional disciplinary boundaries to include anything that clings to us, whether it be Werner Heisenberg's letters or an episode of *Breaking Bad*.

Although instrumental to every humanities project, the wall has a brutally short lifespan. The writer strives to reassert control over its borders and boundaries by whittling down its undisciplined excesses; indeed, training to be a scholar is in large part learning to compress and contain the wall's licentious sprawl. We shorten our focus to a single period, place, or author; excise those fragments that fall outside the increasingly narrow range of our "expertise"; and briskly sever any loose ends that refuse to be tied. These regulatory measures help align our work with the temporal, geographic, and aesthetic boundaries of our disciplinary arbiters: the journals and university presses that publish our work, the departments that hire and tenure us. In an increasingly tight academic marketplace, where the qualified scholars, articles, and projects far outnumber the available positions, deviation from the standard model can seem like risky business indeed.

Even as entrenched structures dictate compression and containment in scholarly writing, the open networks of the web have enabled a publication model based on public sharing and collaboration, spurring a turn to process across the humanities. It has become normal for scholars of all fields to share their incipient, in-progress research on blogs and wikis, and look to the comments sections for peer review. On a larger scale, these moves toward a collaborative process of knowledge-making are visible in the editing policies of Wikipedia; in Femtechnet's Distributed Open Content Course (DOCC), an open repository for course materials; and in new open access imprints like the Dead Letter Office of Punctum Books, which publishes abandoned scholarly projects (to name just a few examples among many). This turn to process has put pressure on the gatekeeping mechanisms described above, as many scholars yearn for a less rigid publishing model that foments the networked creativity of the wall.

Advocating for the transformative effect of a process-oriented model of digital publication, this short paper asks: how can digital humanities not only embrace process rhetorically, but in fact accrete tangible value to the more piecemeal, contingent aspects of knowledge creation? How can we make it the wall's scholarly sprawl "count" within systems that still rely on the trimmed and trussed-up products of research? How can we not only laud conceptually but help to build materially critical practices that eschew disciplinary (and disciplining) boundaries in favor of openings and traversals?

After a brief survey of existing digital journals and other publishing initiatives, including *Hyperrhiz*, Scalar, and Electric Press, we turn to our own incipient venture, titled *thresholds*. *thresholds* is a web-based digital publishing platform for creative scholarship, stitched together from existing digital humanities tools. By sketching the primary design features of *thresholds* – both their theoretical motivations and technical solutions, described in brief below – this short paper argues for a capacious digital publishing model that negotiates, without dissolving, the shifting edges between reading and writing, process and product, the fragment and the collective.

Design Features

The primary design feature of *thresholds* is the split screen. On the webpage's virtual verso are short critical essays that exceed disciplinary boundaries, whether it be in content, style, or approach. We solicit work that a traditional academic journal may deem unfinished, unseemly, or otherwise unbound, but which discovers precisely in its unboundedness new and oblique critical perspectives. Along with her essay, the author submits the textual, visual, and audible fragments that provoked and surreptitiously steered her work. These are published on the right side of the screen and scroll in tandem with the corresponding essay. These scraps are not explicitly harnessed to the work's main body, but instead lie beside it to create provocative juxtapositions; it is left to the reader to forge lines of connection between recto and verso.

Reinforcing its commitment to process and material form, *thresholds* further provides a digital toolkit for readerly making. These tools assign names and haptic functions to those critical traversals that a reader makes through and against a text. As the author's fragments scroll up the right-hand side of the screen, the reader can anchor a piece, holding it in place for future reference, or join one scrap to another to generate new patterns and co-movements. She can also import new material, either by copying text over from the essay on the verso or by composing additional fragments that leak new texts, artists, or ideas into the system. At the end of any given reading session, then, the reader will have generated her own "wall," plucking, amassing, and recomposing the author's fragments to create her own annotative assemblage.

At any time, the reader can capture and conserve the "constellation" that she has produced—that is, the current arrangement of the fragments that she has chosen to lock and join together. Although every user has access to the same firmament of texts that cycle through *thresholds*, each constellation will be singular; their unique spatial architecture will attest to the creative and critical value in visualizing the relations between fragments and texts, readers and authors, and readers and texts. Readers who choose to publically share their work will be able to see how their own creation fits into a galaxy of all other users' constellations, mapping their own choices against that of a collective readership. By enabling the reader to place herself in relation to both the author's text *and* all other readers of the site, *thresholds* models criticism as an intimate yet communal activity that inheres in the delicate links we build in the spaces between each other, as much as between the texts themselves.

To ensure that this intervention is not only conceptually provocative but also formally useful, *thresholds* endows each fragment with a flexible markup language. Readers can download their constellations, receiving a file listing all texts, objects, and art cited therein. This file can then be imported into citation software or shared with others. This underlying information architecture, not immediately present to visitors but baked into the structure of the site, plugs the swirl of scraps that make up any given constellation into the existing citational infrastructure of the humanities. In so doing, it allows *thresholds* to negotiate the gap between that which is in-progress and incomplete within our reading practices—the stray underline, the forgotten marginal note—and more formalized and prescriptive methods for incorporating others' work into our own. There is a place, *thresholds* implicitly argues, for the fragmentary in our collecting and collective practices; for the wall's sprawl within the more regimented systems that order our work.

Bibliography

Benjamin, W. (1999).*The Arcades Project*, trans. Howard Eiland and Kevin McLaughlin. Cambridge: Harvard University Press.

Deleuze, G. (1988). *L'abécédaire de Gilles Deleuze, an interview with Gilles Deleuze, directed by Claire Parnet.*

Farina, L. (2014). Sticking Together. In Cohen, J. J., Joy, E. A., and Seaman, M. (eds), *Burn After Reading/The Future We Want.* Brooklyn: Punctum Books, pp. 31-38.

Adding Semantics To Comics Using A Crowdsourcing Approach

Mihnea Tufis
mihnea.tufis@gmail.com
Pierre and Marie Curie University, Paris 6 (UPMC), France

Introduction

With over 85 million print units sold in 2014 for the top 300 comic book titles only, the comics industry is reaching a new high for the first time since 2007 (before the economical crisis). And this doesn't include the increasingly popular graphic novels or the increasingly more accessible digital comics. The resurgence of comics and the establishment of the graphic novel as a literary genre prompted Humanities scholars to turn their attention on comics as a medium. In this paper, we address the difficulty of creating a digitized corpus by using a crowdsourced approach for annotating comic books. The resulting XML-based encodings could assist not only researchers, but publishers and collection curators equally.

Motivation

Our approach should provide Digital Humanities (DH) scholars with a (currently missing) structured, annotated corpus; this should enable or speed up research related to the comics and sequential art theory: identifying the rhythm of the narration based on the shape, size or number of panels and its relation to the depicted action, investigations about the style of comics authors, historical periods, cultural movements etc.

Curators and collectors (professional or amateur) of physical or online comics collections would be provided with a structured content which could be more easily integrated within their collections or databases. This may assist them into enlarging public or private databases of characters or comics series and enable the creation of artefacts such as comic books dictionaries, search indices and dictionaries of onomatopoeia. A certain number of projects are already in place and could greatly benefit from the creation of comic books annotations. We mention here the Grand Comics Database[1] (an online database of printed comics), Comic Book Database[2], Digital Comics Museum[3] (a collection of scanned public domain comics from the Golden Age) or the Catalogue of the *Cité Internationale de la Bande Dessinée et de l'Image*[4] from Angoulême.

From a publishing perspective, current standard specifications related to digital comics, such as EPUB's Region Based Navigation (Conboy et al., 2014) and Metadata Structural Vocabulary for Comics (Ichikawa et al., 2014) are taking care exclusively of the presentation layer (i.e. rendering a publication on a screen device). But the artistic nature of comics and the great potential digital comics have already showcased allow us to go beyond simple content presentation. We believe that the data we are collecting will allow publishers and digital comics authors to create better, enhanced content and in the end a superior reading experience.

Crowdsourcing Annotations for Comics

Participants to our crowdsourcing experiment (Azavea, 2014; Sharma, 2010) are digital comics readers. Previous research has identified expertise sharing, belonging to a community and helping with a research project as strong motivating factors for crowdsourcing participants (Dunn et al., 2012). In addition, our industrial partner will incentivize participants with product vouchers for their platform[5].

The tasks we propose are organized around a set of questions regarding a series of comics-related topics (Eisner, 1985; Ichikawa et al., 2014) on which computer algorithms are not performing well enough: complex page layouts, identification of narration elements (characters, places, events, objects), stylistic elements (balloon shapes, onomatopoeia, movement lines).

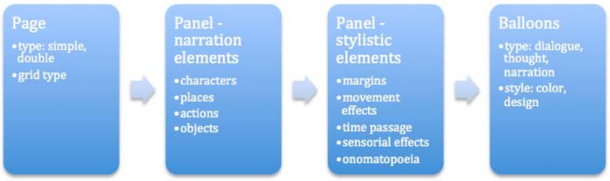


Figure 1. The 4 key annotation themes

We aggregate the answers (Feng et al., 2014; Snow et al., 2008) taking into account the reliability of an annotator in a given context (task difficulty, user experience with the task, type of question) and the agreement between the annotators (Nowak et al., 2010). A quality score is thus generated for each annotation, with the best of them being selected as solutions.

We subsequently are able to generate the ComicsML encodings (Walsh, 2012). This XML derived format is particularly useful since it's based on the already widespread Text Encoding Initiative (TEI), allowing for declarations of page structure and composition, panels, characters, text (in all the varieties hosted by the comics medium: different types of balloons, diegetic text, onomatopoeia), events and even panel-to-panel transitions.

Page structure

The annotators are presented with a simple interface (Fig. 2) in which they will have to make a choice between a set of suggested grid layouts. These layouts are the output of applying the automatic frame extraction algorithm developed by (Rigaud et al., 2011). Alternatively, for complex page layouts, they will be asked to draw the page layout themselves.

Character identification

At this step, we ask the "crowd" to simply enumerate all the characters they can identify in the current page. Characters are identified by reading their names in the text, recognising them from experience or simply giving a general statement about the character (e.g., " *masked man*" may be referring to *Batman*). Using state of the art symbolic learning algorithms we fusion generic and specific information (e.g. if *Batman* and "*masked man*" both have high quality scores in the same context, they will both denote the same concept and will be considered as valid annotations).

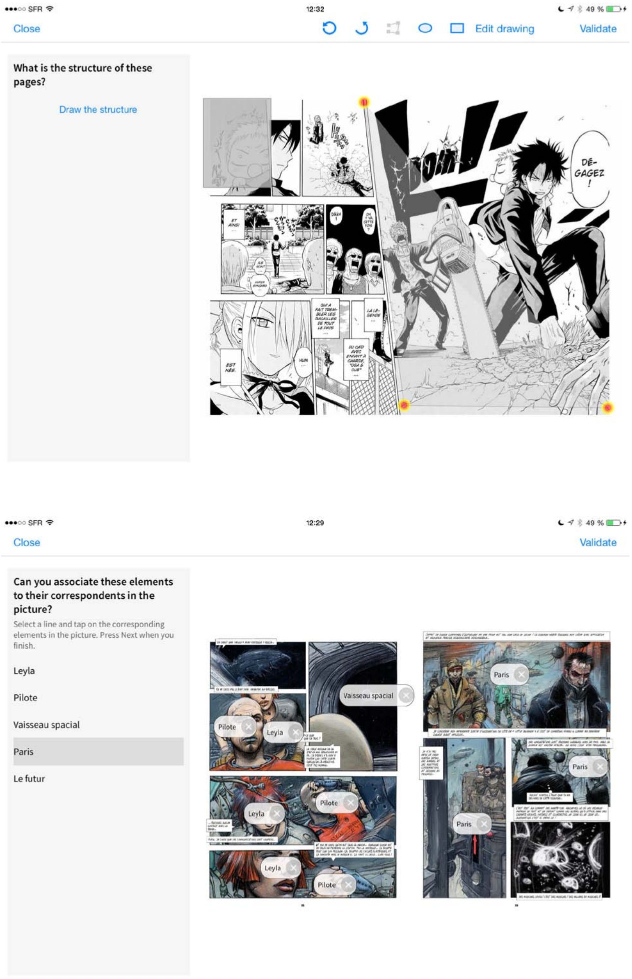
Places identification

We ask the annotators to simply enumerate all the places they can recognise on the current page. We are particularly looking for named places (e.g., *Gotham City, NY, planet Mars*), but will also ask the annotator to mark any generic place that he might consider important for the scenes in the page (e.g., “ *the interior of a bank*“ [in case of a robbery], “*inside a space ship*” [in case of a battle in space]). These are exactly the kind of very specific annotation tasks for which state of the art image recognition algorithms are expected to fail.

Events identification

This is yet another highly specific recognition task. Annotators are asked to describe the most important events occurring in the page. The solutions generated at this step, together with the annotations obtained in the previous steps will be used to further build the ComicsML encoding of the page.

Comics scholar Scott McCloud stresses the role of elipsis (“the blood in the gutters“ – the space between two panels) as an artistic mean for authors to engage their readers, and describes a typology of these transition spaces (McCloud, 1993). ComicsML allows us to declare such transitions through the #ana attribute of the cbml:panel element, giving us the possibility to investigate their use and their distribution across different cultural spaces (France/Belgium, Japan/Korea, USA).



What is the structure of these pages?

Can you associate these elements to their correspondents in the picture?

Select a line and tap on the corresponding elements in the picture. Press Next when you finish.

Figure 2. Annotation interface (drawing page structure-left, narration elements-right)

Non-visual cues

Comics are a special medium, making use of the visual to depict all other non-visual senses, with the help of different drawing “tricks”, such as:

- Smoke coming out of a cigarette may engage the reader’s smelling sense
- Onomatopoeia form a particular language of their own; comics and especially manga authors have proven great creativity when it comes to expressing different sounds via stylised text (e.g., “ *POW!*” – punch, “ *BAM!*” – gunshot)
- Horizontal lines around a car suggest the car is moving at high speed, while around a ball, they express the ball’s movement.

Researchers could study, for instance, the drawing style of an author and his use of non-visual cues, and go as far as creating onomatopoeia dictionaries for comics (to our knowledge, such dictionaries already exist for manga, but not for American nor European comics).

At the end of this stage, we should be able to generate a reasonably complete ComicsML encoding of the current page (see Fig. 3).

```
<cbml:panel
    n="2"
    characters="#pilot #leyla"
    ana="#subject-to-subject"
    xml:id="case_002"
    xmlns:cbml="http://www.cbml.org/ns/1.0">
        <cbml:balloon xml:id="boule_002" type="speech" who="#pilot">
            You should reconsider. Look above!
        </cbml:balloon>
        <cbml:balloon type="speech" who="#leyla">
            What is this thing?
        </cbml:balloon>
</cbml:panel>
```

Figure 3. A fragment of the ComicsML encoding for the page presented above

Conclusions

We have presented the outline of our crowdsourced annotation system for comics, as well as details of how we have designed our tasks, having in mind three main aspects: the limits of current digital comic book formats, the specifications behind the ComicsML metadata schema and theoretical principles of comics as a medium (Eisner, 1985). Last, we have presented the way in which the collected results are merged into the final ComicsML encoding and have briefly discussed some potential applications.

Bibliography

Azavea, SciStarter. (2014). *Citizen Science Data Factory, Part 1: Data Collection Platform Evaluation*.

Dunn, S. and Hedges, M. (2012). Engaging the Crowd with Humanities Research. *Crowd-Sourcing Scoping Study*. Centre for e-Research, Dept. of Digital Humanities – King’s College, London.

Conboy, G., Duga, B., Gardeur, H., Kanai, T., Kopp, M., Kroupa, B., Lester, J., Garrish, M., Murata, M. and O'Connor E. (2014). *EPUB Region Based Navigation 1.0*. http://www.idpf.org/epub/renditions/region-nav/ (accessed 5 March 2016).

Eisner, W. (1985).*Comics and Sequential Art: Principles and Practices From the Legendary Cartoonist*. Norton&Company, USA&U.K.

Feng, D., Sveva, B. and Zajac, R. (2009). Acquiring High Quality Non-Expert Knowledge from On-demand Workforce. *People's Web Meets NLP-2009*, ACL (2009), 51-56.

Ichikawa, D., Kasdorf, B, Kopp, M. and Kroupa, B. (2014). *EPUB Region Based Navigation Markup Guide 1.0*. http://www.idpf.org/epub/guides/region-nav-markup/ (accessed 5 March 2016).

McCloud, S. (1993). *Understanding Comics – The Invisible Art*, Harper Collins, USA.

Nowak, S. and Ruger, S. (2010) How reliable are annotations via crowdsourcing? A study about inter-annotator agreement for multi-label image annotation. In *Proc. MIR-2010*, ACM, 557-566.

Rigaud, C., Tsopze, N., Burie, J.-C. and Ogier, J.-M. (2011). Robust text and frame extraction from comic books. *GREC-2011*, Springer, 129-138.

Sharma, A. (2010). Crowdsourcing Critical Success Factor Model: Strategies to harness the collective intelligence of the crowd. Working paper.

Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y. (2008). Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *EMNLP-2008*, ACM, 254-263.

Walsh, J.A. (2012). Comic Book Markup Language: an Introduction and Rationale. *DHQ-6*, 1. http://www.digitalhumanities.org/dhq/vol/6/1/000117/000117.html (accessed 5 March 2016).

Notes

[1] http://www.comics.org/
[2] http://comicbookdb.com/
[3] http://digitalcomicmuseum.com/
[4] http://www.citebd.org/
[5] Actialuna – Sequencity: https://www.sequencity.com

Mapping Multilingual Responses To Famine And Dearth In The Early Modern Landscapes Of India And Britain

Charlotte Tupman
c.tupman2@exeter.ac.uk
University of Exeter, United Kingdom

Richard Holding
r.j.holding@exeter.ac.uk
University of Exeter, United Kingdom

Hannah Petrie
h.petrie@exeter.ac.uk
University of Exeter, United Kingdom

Gary Stringer
g.b.stringer@exeter.ac.uk
University of Exeter, United Kingdom

Famine, as John Walter writes, was "…a recurring reality and ever-present fear" in the early modern period (Walter, 2015). The AHRC-funded project *Famine and Dearth in India and Britain, 1550-1800: Connected Cultural Histories of Food Security* (http://humanities.exeter.ac.uk/english/research/projects/famine/) examines the practices, discourses and literary modes through which societies in early modern India and Britain articulated their concerns about the availability and distribution of food (Mukherjee, 2014). A collaboration between the University of Exeter and Jadavpur University, Calcutta, the project draws upon a large body of texts written in languages including English, Latin, Persian, Bengali and Hindi for evidence about cultural responses to landscapes of famine and dearth. Its aim is to produce a digital resource that includes encoded extracts from the source materials and maps that reflect the variety and scope of identified responses to the landscapes encountered in the sources.

The application of digital methods to large corpora of thematic texts presents both opportunities and challenges, which will be examined in this work-in-progress paper. The dataset is highly diverse, not only in terms of the languages in which these texts are written but also in the types of documents that form our body of evidence. The source materials include chronicle histories, gazetteers, official correspondence, legislation, pamphlets, periodicals, plays, poetry, surveys and prose (fiction and non-fiction), and the project will publish excerpts from each of these categories. Our markup priorities (beyond the basic structure of the document) lie in how to encode the many themes surrounding famine and dearth that are present in our texts, which we need to extract in order to address the research questions of the project. We can use natural language processing tools to help us identify names and places, for instance (at least for some of the languages), but the range of terms used in these texts to describe the features of the landscape, the people, and the food situation are extremely varied. Inevitably there are also subtleties in the ways in which particular concepts are represented in the various languages of our source materials, as well as added complications such as variant spellings. The project uses the open source software GATE (General Architecture for Text Engineering, http://gate.ac.uk/) to identify names and places in the texts written in English, and we would like to apply a similar process to the texts written in other languages. It is likely that we will create some of the gazetteers from scratch, in which case we would aim to make these available as part of the project's outputs.

One of the most interesting challenges is in how to map the resulting data meaningfully. Exploring responses to landscapes of scarcity is a key research question of the project, and descriptions of such responses feature very heavily in some of our texts, particularly in the travel writings (see McRae, 2009 for a discussion of some of themes in travel writing of this period). The works of Peter Mundy, for instance, are full of rich descriptions of the places he visited, including very personal observations of the circumstances in which he found himself (Carnac Temple, 1914; Pritchard, 2011). We are particularly keen to place our work in the context of some of the projects that have taken place during the last decade on mapping emotional responses to landscapes at other periods and in different geographical areas. While many of the recent projects on emotional cartography use wearable technologies to measure and record responses to the landscape, such as Christian Nold's 2006 Greenwich Emotion Map (an art project combining annotations with measurements of skin responses at different stages of a walk through the Greenwich area of London: http://www.emotionmap.net/background.htm), we see potential in learning from, and building upon, their approaches to visualising the resulting data (Nold, 2009). Projects such as Kurt Jensen's representation of Sterne's *A Sentimental Journey*, built using Neatline (http://neatline.org/), have also suggested possible directions for representing some of the travel writings (http://enec3120.neatline-uva.org/neatline/show/a-sentimental-journey), and our recent workshop on food security has helped to clarify the relevant user requirements (http://foodsecurity.exeter.ac.uk/). However the question of how to integrate such a wide variety of sources and languages into useful and meaningful maps remains one of the most interesting and challenging aspects of the project. As such, it will be a key focus of our paper, and we anticipate that presenting the results of our experiments with this data could be helpful for other projects that are grappling with similar issues, potentially in very different subject areas.

Bibliography

Carnac Temple, R. (ed.). (1914). *The Travels of Peter Mundy in Europe and Asia, 1608-1667*. London: printed for the Hakluyt Society.

McRae, A. (2009). *Literature and Domestic Travel in Early Modern* England. Cambridge: Cambridge University Press.

Mukherjee, A. (2014). *Penury Into Plenty: Dearth and the Making of Knowledge in Early Modern England*. London and New York: Routledge.

Nold, C. (ed.). (2009). *Emotional Cartography – Technologies of the Self.* Available at http://emotionalcartography.net/ Accessed 5th March 2016.

Pritchard, R.E. (2011). *Peter Mundy, Merchant Adventurer*. Oxford: Bodleian Library, University of Oxford.

Walter, J. (2015). Abstract for "Poverty without patience? The politics of dearth and scarcity in early modern England", presented at the workshop 'Food Security: Past and Present', Oxford, 3-4 September 2015.

A lesson in applied minimalism: adopting the TEI processing model

Magdalena Turska
tuurma@gmail.com
University of Oxford, United Kingdom

James Cummings
james.cummings@it.ox.ac.uk
University of Oxford, United Kingdom

The Guidelines of the Text Encoding Initiative Consortium (TEI) have been used throughout numerous disciplines producing huge numbers of TEI collections. These digital texts are most often transformed for display as websites and camera-ready copies. TEI Simple (Rahtz et al, 2014) project was the first one to propose more prescriptive approach providing the baseline rules of processing TEI into various publication formats, while offering the possibility of building customized processing models within the same infrastructure. For the first time in history of TEI there exists a sound recommendation for default processing scheme, which significantly lowers the barriers for entry-level TEI users and enables better integration with editing and publication tools. The TEI Simple project was a Mellon-funded collaboration between the TEI Consortium, Northwestern University, the University of Nebraska at Lincoln, and the University of Oxford.

The new (on track for acceptance by early 2016) TEI method for documenting processing models gives editors and TEI customisers a method for high level recording of processing intentions in a machine-processable but implementation agnostic manner. Nevertheless, the processing model is a new proposal and needs to be extensively tested before announcing it a success. As the TEI Technical Council works to integrate the TEI processing model extensions created by the TEI Simple project, we endeavour to employ it on real world projects, both of which have been running for a significant number of years and have already produced vast collections of material: historical documents of the US Department of State, Office of Historian (http://history.state.gov/) and the corpus of Ioannes Dantiscus' correspondence (http://dantiscus.al.uw.edu.pl/). The Office of the Historian publishes a large collection of archival documents of state, especially those appartaining to foreign relations. The Dantiscus project spans over ten thousand original sources from the early sixteenth century including correspondence, poetry, and diplomatic documents. This makes it a good test case for implementation of the TEI Processing Model because it is far beyond the scope of the original TEI Simple sample collections. Having the material previously published with custom-built XQuery/XSLT packages means that we are in a position to compare the results of using an approach based on the processing model with the previous one in terms of the quality and consistency of final presentation but also in more quantitative ways like the size of necessary code base, development time and ease of the long-term maintenance.

The first challenge is, obviously, rephrasing the transformations previously formulated in XQuery/XSLT using ODD meta-language extensions proposed by TEI Simple project. Preliminary results are very encouraging even though, as expected, it became necessary to extend the behaviours library to accommodate some specific needs. From the developer's perspective it is immediately clear that using the TEI processing model brings substantial advantages in development time and results in much leaner and cleaner code to maintain. For the Office of Historian project figures suggest code reduction by at least two-thirds in size. Numbers are even more impressive realizing that the resulting ODD file is not only smaller, but much less dense, consisting mostly of formulaic <model> expressions that make it easier to read, understand and maintain, even by less skilled developers.

To a lesser extent, but it is still interesting to see if, thanks to the additional layer of abstraction that processing model brings to the table, the editors can become more independent from developers in tweaking the processing rules. This heavily depends on the personal predilections of the editor, but again, in cases where editors are already deeply involved in the decisions about encoding on the level of XML markup and do have some fluency in XPath and/or CSS our results show that it is perfectly reasonable to expect them to tailor the existing high-level processing models to fit their specific needs in a majority of cases. We will also investigate the effect of incorporating the Processing Model into eXist-db native database and application framework (Meier et al, 2016) environment in terms of easening the learning curve, for the non-technical users in particular.

The processing model at the time of writing this paper proposal is not a mature technology yet, in the sense that it still lacks the critical mass of its practitioners as well as formal acceptance by the TEI Technical Council (although this will have been integrated into the TEI infrastructure by the time of DH2016). This presentation aims to present both challenges and open questions as well as already demonstrated advantages of applying this technology. It will draw on the evidence from early adopters available by the time of DH2016. It is not only the quantitative measures of improvements in technical implementations that will be reported on, but the variation in methodologies employed by the test projects and others.

Bibliography

Meier, W. and Turska, M.(2016). TEI Processing Model Toolbox Documentation,http://showcases.exist-db.org/exist/apps/tei-simple/doc/documentation.xml?odd=documentation.odd(accessed 5 March 2015)

Rahtz, S., Mueller, M., Pytlik-Zillig, B., Turska, M. and Cummings, J.(2015). TEI Simple Processing Model Specification,http://htmlpreview.github.io/?https://github.com/TEIC/TEI-Simple/blob/master/tei-pm.html(accessed 5 March 2015)

Exploring Networks Of Confidentiality And Secrecy In Early Modern Transconfessional Correspondences

Ingeborg van Vugt
ingeborgvanvugt@hotmail.com
Scuola Normale Superiore di Pisa, Italy

With this contribution, I would like to discuss how multi-layered visualizations of epistolary networks can contribute to a better understanding of the circulation of illegal literature and confidential ideas between Catholic Tuscany and the Calvinist Dutch Republic. It questions how intellectual exchanges between these two regions maintained a balance between, on the one hand, the necessity to distribute (prohibited) books and to express controversial ideas and, on the other, social control and the need to avoid the objections of powerful political and religious institutions and individuals. This comparative analysis allows for a sharper focus on the differences and similarities on how intellectuals capitalized on opportunities in the social and religious structures to which they were connected. Indeed, they had to deal with the many tensions between the oppressive catholic environment of the court of Cosimo III and the Dutch Republic, already well known for its relative tolerance and freedom of printing (e.g. Touber, 2014).

These personal and societal conflicts forced scholars and booksellers to take strategic measures of secrecy and confidentiality, which in turn depended on what Mauelshagen (2003) also called "networks of trust". If we pose the question how epistolary networks evolved, understanding changing relationships between people, one might provide insights into aspects of confidentiality. For instance, as relationships grew friendlier, correspondence grew in confidence and trust, while on the other hand one did not correspond with adversaries (Heuvel, et al. 2014). Other examples include studies of Lux and Cook (1998), who claimed that the success of the Dutch Republic depended on what Granovetter (1973) also called "weak ties" instead of central hubs. This implies the importance of intermediaries between communities for the faster distribution of ideas: if you have a particularly close friend from another community, you are more likely to introduce him to your other close friends whom you know that will trust you (Barabási, 2009: 55).

If we wish to verify these statements and to understand how intellectuals were able to overcome these confessional and social barriers, the analysis of multi-layered networks of correspondences provides a very interesting addition to archival research. Or in other words, the combination of methods for network analysis for distant reading of large sets of letters with close reading devoted to detect the role of secrecy and confidentiality in epistolary exchanges strengthens historical research. This means that qualitative analysis will uncover how social relations are represented and constructed, sometimes reinforced and sometimes transformed, which is enriched by traditional hermeneutic methods to focus on specific religious and personal features that have influenced those dynamics.

It is important to consider that full data integration, in particular when dealing with early modern letters, is impossible for reasons of incompleteness, complexity and uncertainty in data. Therefore the focus should not be on analytical and statistical methods of network representations alone, but on approaches that allows us to handle, inquire and interpret these complex historical data. We do not need just networks as static representations, but also networks as interactive interfaces (Heuvel, et al. 2016). To this end, the software tool Nodegoat is used to bring together, explore and contextualise these epistolary exchanges. Nodegoat, differently from network visual-analytical tools such as Gephi, is built around data entry, management and curation processes (http://mnn.nodegoat.net/viewer). It enables us to explore and to combine historical networks in various configurations, involving a diverse set of actors. This means that not only persons (scholars and booksellers) constitute relationships but also textual objects, like books with their dedicatees and introduction letters. For instance, the overlay of networks highlights those artefacts that played an important role in the establishment of contacts. If we look at their intersection with their function in the network, new opportunities may rise about how to link book dedications to strategies adopted by scholars. The importance of textual objects as participants in networks has also been stressed by Latour (2005). From the correspondence of the Florentine librarian Antonio Magliabechi, for example, it turned out that the Dutch microbiologist Antoni van Leeuwenhoek (1632–1723) chose to dedicate his work the *Arcana Naturae Detecta* to Magliabechi. By using this strategy, Leeuwenhoek was able to benefit from Magliabechi's extensive network for the distribution of his work in Italy. Overlaying different networks sheds light on the role of the *Arcana* in the network of Magliabechi (fig. 1 and 2). For instance, the image illustrates those correspondences in which Magliabechi mentioned the publication of Leeuwenhoek, showing in this way the diffusion of his publication over time.

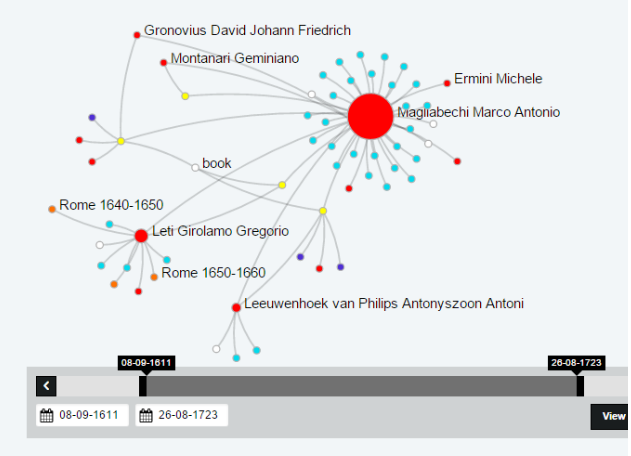
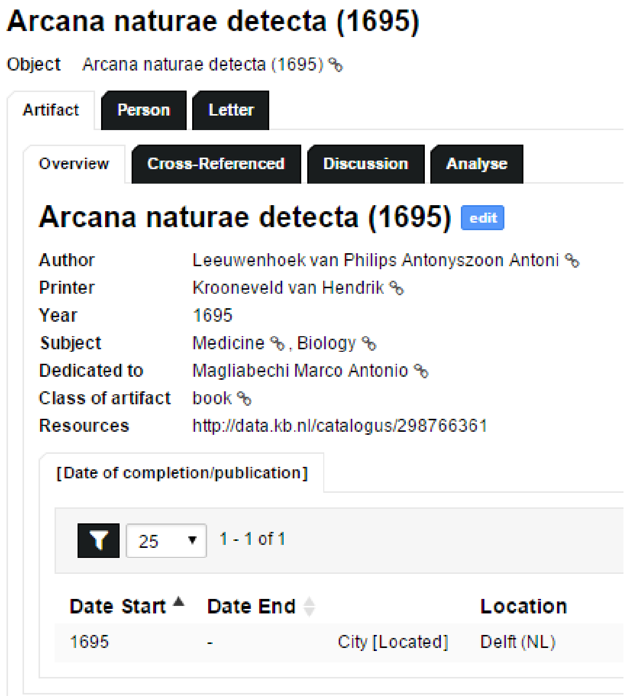


Fig. 1 A visualization of networks around the *Arcana Naturae Detecta* of Leeuwenhoek (the right yellow node) and Magliabechi. Magliabechi, at the centre of this visualisation, is surrounded by other dedicatees (represented in red), their accompanying books (yellow) and the letters in which the *Arcana* is mentioned (light blue). Hovering over the nodes and ties opens an overview with the different connections, and specifies the nature of the relationships (see fig. 2)

Moreover, the analysis of relationships between people can provide insight between direct and indirect transfer of confidentiality via intermediaries. In the Mapping Notes and Nodes in Networks project (Álvarez and Heuvel, 2014) it appeared that the overlay of more networks shed light on the evolution of co-citation networks and introduction networks. As correspondents entered networks of epistolary exchange, they did so not in some ideal egalitarian society, where anyone could join simply by writing a letter, but in a world regulated by social norms and rules of etiquette. In short, letters of introduction were often necessary to be admitted into an epistolary network. Following the evolution of introduction letters alongside a citation network and determine whether a shift takes place in the number of intermediates between correspondents (revealing shrinking degrees of separation), could reveal the importance of introductions in the establishment of epistolary networks.



Arcana naturae detecta (1695)

Object Arcana naturae detecta (1695)

Arcana naturae detecta (1695)

Author	Leeuwenhoek van Philips Antonyszoon Antoni
Printer	Krooneveld van Hendrik
Year	1695
Subject	Medicine , Biology
Dedicated to	Magliabechi Marco Antonio
Class of artifact	book
Resources	http://data.kb.nl/catalogus/298766361

[Date of completion/publication]

25 1 - 1 of 1

Date Start	Date End		Location
1695	-	City [Located]	Delft (NL)

Fig. 2 Overview of the *Arcana Naturae Detecta*, linked to the Short Title Catalogue of the National Library of the Netherlands (KB)

Furthermore, Nodegoat is used to bring and contextualise epistolary networks by means of data integration from various data resources such as the Short Title Catalogue Netherlands (STCN) and archival research from archives in the Netherlands and in Italy. The STCN has been queried in order to disambiguate objects and to enhance the interoperability of my data. For example, information on books in epistolary networks can be linked directly to the STCN, which allows me to map the unstructured data in the letters (as titles are often mentioned incomplete) to structured data.

Bibliography

Álvares Francés, L., Heuvel, C. van den (2014). *End report Mapping Notes and Nodes in Networks.* https://www.huygens.knaw.nl/wp-content/uploads/2015/05/EndReportMNN.pdf> (accessed 18 December 2016).

Barabási, A. (2002). *Linked; the New Science of Networks*. Cambridge/Massachusetts: Perseus Publishing.

Granovetter, M. S. (1973). The Strength of Weak Ties, *American Journal of Sociology*, **78**(6): 1360-1380.

Heuvel, C. van den, Vugt, I. van, Kessels, G., Bree, P. van (2016). Deep networks as associative interfaces to historical research, The Future of Historical Network Research (peer reviewed book chapter) in *Ashgate* (submitted for publication).

Heuvel, C. van den et al. (2015), Modeling Confidentiality and Secrecy in Knowledge Exchange Networks of Letters and Drawings in the Early Modern Europe, *Nuncius*, 31: 78–106.

Latour, B. (2005). *Reassembling the Social, An Introduction to Actor-Network-Theory*. Oxford: University Press.

Lux, D and Cook, H. (1998). Closed circles or open networks? Communicating at a distance during the scientific revolution, *History of Science*, **36**(112): 179-211.

Mauelshagen, F. (2003). Networks of Trust and Imagined Community of the Learned, *The Medieval History Journal*, 6(1): 1-32.

Touber, J. J. (2014). Religious interests and scholarly exchange in the Early Enlightement Republic of Letters: Italian and Dutch scholars 1675-1715, *Rivista di Storia della Chiesa in Italia*, 2: 411-436.

Using Big Cultural Data To Understand Diversity And Reciprocity In The Global Flow Of Contemporary Cinema

Deb Verhoeven
deb.verhoeven@deakin.edu.au
Deakin University, Australia

Bronwyn Coate
bronwyn.coate@rmit.edu.au
RMIT University, Australia

Colin Arrowsmith
colin.arrowsmith@rmit.edu.au
RMIT University, Australia

Stuart Palmer
stuart.palmer@deakin.edu.au
Deakin University, Australia

This paper explores the relationships between countries in the exchange of movies and measures the reciprocal nature of these relationships. This investigation represents an innovative way to explore international exchanges of digital cultural content based on global cinema screenings analysed at the national level. Rather than focus on the market dominance of particular cinemas (e.g. the US or Indian cinemas) we examine the relative strength of two-way relationships in order to understand cultural reciprocity in the film industry. The dynamics of shared cultural exchange are explored in terms of the volume of transactions between ‘cinema nations’ expressed in the form of dyadic networks.

The paper is based on the premise that films can be understood as cultural goods that are distributed both between 'territories' or markets and across the globe according to industrially unique spatial patterns and temporal flows. Seeing film diffusion in this way invites us to explore the industrial aspects of movement and location but it also invites reflection on our use of these large datasets. For example, understanding the dynamics of global film exhibition and distribution demands an appreciation of scale and velocity in both the film industry and in a data-driven approach to its study. Data-driven approaches to Cinema Studies are at best an emergent aspect of the discipline (Verhoeven). This paper makes a significant contribution to the development of Cinema Studies by extending a trans-disciplinary, digital humanities approach to critically understanding the dimensions of a global creative industry at scale.

Digitisation and globalisation are full of contradictions in terms of how they impact diversity of screen culture represented by film. On one hand digitisation has facilitated an explosion in the number of films being made and that can be distributed and viewed online. This has had the effect of increasing the diversity of films available to audiences with digital access over the web. On the other hand however, only a relatively small proportion of films produced are released widely into cinemas. This paper seeks to provide insight into diversity at cinema locations that extends beyond the obvious dominance of Hollywood blockbusters. We are interested in drawing attention to equitable reciprocal exchange relationships that exist between nations, even where these may be small in scale, as evidence of alternative practices in the promotion of diversity at the cinema. This enables us to explore relational geographies using dyads in an approach similar to that of Taylor, Hoyler, Pain and Vincigurrra (2013) in their investigation of the connectivity between cities in the services sector. We use dyadic analysis to explore an equitable exchange in film that extends beyond the unilateral to ensure cultural exchange between two nations is assessed as a two-way flow where cultural content from both sides to the dyadic relationships are valued and accounted for.

The data used in this paper is drawn from the Kinomatics Global Showtime dataset (Kinomatics, 2015) which comprises over 330 million individual records of film screenings from across 31,500 venues covering 47 countries, including the US, India, most of Western Europe, Japan, Brazil and Australia for the years 2013 – mid-2015 (see Arrowsmith et al). For this analysis, the kinomatics data is analysed using a variety of methods that draw from a range of disciplinary perspectives including the digital humanities and economics as well as geospatial and computational sciences. Our key tool is Principal Components Analysis that explores dyadic relationships as part of Social Network Analysis (see: Wasserman and Faust, 1999). We apply this approach to 'nations' as they are defined by the kinomatics (cinema screening data aggregated at the national level) and imdb (film production aggregated at the national level) datasources. Further to this, in a selection of countries, we employ a Herfindahl-Hirschman Index (HHI) to consider case study analysis of cultural diversity based on cinema screen count data. We use spatio-temporal visualisations as a way to represent the results and propose insights into the relational geographies of film flow and exchange that are found to exist.

Using dyads to explore international exchanges between nations we are able to consider diversity in relation to films screened at the cinema in terms of the two key dimensions stemming from globalised relations between nations, namely in terms of intensive and extensive international exchanges. The intensive dimension aids understanding of the most important national dyads that dominate cinema screenings that can be seen as the core centres of the globalised market for film, while the extensive dimension is focused upon the multitude of links between nations in a broader globalised market for film screened at the cinema. The analysis of dyadic relationships enables us to move beyond the assumption that the flow of cinema is simply unilateral. Instead we are concerned with the relative strength of exchanges in which a strong reciprocal dyadic relationship is one that has an equal exchange between two nodes, in this case, countries.

In considering diversity using the HHI we focus on the screening of new release features within the case studies of Australia, France and the Republic of Korea in order to analyse the relationship between cinema venue location (capital cities versus regional), venue type (in terms of the number of screens) and film programming allocations between domestic, US and other imported feature films. We find that as an increasing number of films are being released, non-US derived films are struggling in a tight contest for screen time.

Bibliography

Arrowsmith, C., Verhoeven, D., Davidson, A. and Coate, B. (2014). 'Kinomatics: A global study into Cinema Data', *Proceedings of the GSR_3 Conference*, http://kinomatics.com/wp-content/uploads/2014/12/GSR_Kinomatics.pdf, Melbourne

Kinomatics (2015), www.kinomatics.com

Taylor, P., Hoyler, M., Pain, K. and Vinciguerra, S. (2013) 'Extensive and intensive globalisations: Explicating the low connectivity puzzle of U.S. cities using a city dyad analysis', *Journal of Urban Affairs*, Vol 36, issue 5, doi:10.1111/juaf.12077, 1-14.

Verhoeven, D. (2012) 'New Cinema History and the Computational Turn', Beyond Art, Beyond Humanities, Beyond Technology: A New Creativity", *World Congress of Communication and the Arts Conference Proceedings*, University of Minho, http://kinomatics.com/wp-content/uploads/2013/10/Verhoeven_ComputationalTurn.pdf, Portugal

Wasserman, S. and Faust, K. (1999) *Social Network Analysis: Methods and Applications*. Cambridge University Press: New York.

"Digital" in practice: survey of Russian historians' research practices

Andrei Volodin
volodin@hist.msu.ru
Moscow State University, Russian Federation

Realm of Digital history

"Digital history" became a professional realm, because new ways to store, process, and study information entered the everyday historical research practices in Russia (Rosenzweig, 2010). Obviously, the "digitization" introduces new approach to the craft of the historian, when digitization became the indispensable stage of many studies (Volodin, 2015). Interdisciplinary framework of "digital history" is often placed in the context of a broader movement of the "digital humanities" (Schreibman et al., 2004; Schreibman et al., 2016). However, it should be emphasized that historical research has lots of specific features, and "digital turn" makes a significant impact on the research practices (Weller, 2013).

Historical research is a kind of reconstruction of the past on the basis of the extant monuments and documents, "remains" and "legends". This reconstruction is based on three pillars: heuristics, criticism and interpretation. These basic practices today meet the challenges of the "digital age". It's important to mention that the "digital turn" in the profession of historical research was one of the four main themes of the International Congress of Historical Sciences in 2015. It seems that time has come to investigate the specificity of research practices and digital tools for professional historians.

Survey of "digital" in practice

The research is based on questionnaires and interviews of historians in Russia in 2015 and 2016 (178 respondents). In questionnaires and interviews I tried to ask not only "geeks" among historians, but also "pure" historians whom "digital turn" also touched (cf.: DARIAH Survey, 2015; Schreibman and Hanlon, 2010; Trinkle, 1999). First questions ask respondents to name main "digital practices" that they consider important. Such start helps to create "folksonomy" of practices and tools that historians define for themselves as "digital" (e.g. it's not obvious whether text-editing is considered as digital practice, or email as research tool?). Then I started to ask questions about different fields of digital practices as we hypothetically classified in several groups: access, digitization, editing, mark-up, visualization, modeling/simulation, and others. Each group (or type of practice) includes different techniques and tools. We define techniques as more general term to call research procedure, when tools are linked to particular software or platform.

History in the "digital turn"

The instancy of the research is associated with a new stage in the development of information and digital technologies in history. "Digitizing" is not just a technical procedure now, but it becomes a preparation for further computerized analysis of historical sources, using different tools for visualization and analysis of historical information.

The "digital turn" in history requires a monographic study as well as a full systematization of historians' research practices. We don't know any systematic attempt to generalize the results achieved by "digital history". This research project is designed to systematize the research practices (both visualization and analysis of historical data). This study tries to "catch" the development of contemporary digital practices, and we hope to make classification that will summarize achievement in the field of "digital history". Summarizing the answers, it's possible to draw several valid observations: one insight and three dichotomies.

Insight: Digital is "Invisible"

It looks that a prediction which was vividly pronounced in "A New Companion to the Digital Humanities" that "a decade or two from now, the modifier "digital" will have come to seem pleonastic when applied to the humanities" is much closer. Most of answers were concerned with some concrete experience of usage of digital tools, mentioning that they are so "natural" or "spontaneous". The time has come when historians are really in the majority of cases are so-called "digital natives" (even if not in age, but at heart).

The usual remark on state-of-art in historical digital practices was that so-called new ICT (Information and communications technology) became usual, familiar, and mostly routine ICT. The increase of computing power broadens digital practices form tables and databases to full-text search in enormous online collections, and then to visualizations from GIS to 3D (cf.: Gregory, 2014).

Heuristics dichotomy: Analog versus Digital

Digital in Russian literally means numeric (like chiffre or numerique in French, or Zíffern in German) without any link to fingers, but strongly connected to quantitative research and computing. That's why for main historians "digital history" strongly refers to quantitative history or cliometrics. History as one of "traditionally print-based disciplines" (Hayles, 2012, 1) is deeply connected to studies of analog primary sources. This dichotomy has two implications: on the one hand, historians usually see no principle differences between analog original and digital copy, but, on the other hands, they mainly prefer to study analog archives, because it shows their professional skills of access to rare documents.

Criticism dichotomy: Capta versus Data

The “digitization” became the usual and common practice of collecting information. But it's important to mention that for many respondents digitization is almost the same practice as “writing out” or “making notes”, not planning the future possibilities of processing this “capta”.

Distinction between “capta” and “data” in digital history became indicative. If capta is “taken” actively while data is assumed to be a “given” (Drucker, 2011), in digital history capta explains the “catch” of the researcher in terms of his/her sampling of primary sources as well as a critical attitude to the choice of the sources.

Interpretation dichotomy: Close versus Distant

The distinction between “close reading” and “distant reading” came to digital history from computer linguistics (Moretti). This dichotomy explains the conscious choice between brains or processors in solving different research problems. And main present question (or prospect) is that to what extent main practices of history can be robotized? This vision of the future is usually mixed with vague hope that at least interpretation will remain in the researches’ hands and heads. It also brings up an issue of possibilities of digital infrastructure for analytical level.

The aim of the lasting research of digital practices is to create a list of solutions and techniques developed in real research practice, and thus to disseminate such solutions among historians and researchers of the past. The classifier of digital methods and tools will help to choose the most appropriate solution for particular research aims.

Bibliography

DARIAH Survey (2015). “DARIAH Survey on Digital Practices in the Arts and Humanities”. https://dariahre.hypotheses.org/285

Drucker J. (2011) “Humanities Approaches to Graphical Display” in DHQ. 5.1. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Gregory I. (2014) Challenges and opportunities for digital history. Front. Digit. Humanit. 1:1. 10.3389/fdigh.2014.00001

Hayles N.K. (2012) How We Think: Digital Media and Contemporary Technogenesis. University of Chicago Press.

Moretti F. (2013) Distant Reading. Verso.

Rosenzweig R., Grafton A. (2010) Clio Wired: The Future of the Past in the Digital Age. Columbia University Press.

Schreibman S., Siemens R., Unsworth J. (2004) A Companion to Digital Humanities. Blackwell Publishing. http://www.digitalhumanities.org/companion/

Schreibman S., Siemens R., Unsworth J. (2016) A Companion to Digital Humanities. Wiley-Blackwell.

Schreibman S., Hanlon A.M. (2010) “Determining Value for Digital Humanities Tools: Report on a Survey of Tool Developers" in DHQ (4/2). URL: http://digitalhumanities.org/dhq/vol/4/2/000083/000083.html

Trinkle D.A. (1999) “History and the Computer Revolutions A Survey of Current Practices” in Journal of the Association for History and Computing. http://hdl.handle.net/2027/spo.3310410.0002.107

Volodin A. (2015) «Cifrovaja istorija»: remeslo istorika v cifrovuju epokhu [Digital history: the craft of historian in the digital age] // Elektronnyj nauchno-obrazovatel'nyj zhurnal «Istorija», 2015. № 8 (41). 10.18254/S0001228-9-1

Weller T. (2013) History in the digital age. London; New York: Routledge, 2013.

The School of Salamanca on the Semantic Web

Andreas Wagner
andreas.wagner@em.uni-frankfurt.de
Akademie der Wissenschaften und der Literatur | Mainz, Germany

Ingo Caesar
caesar@rg.mpg.de
Akademie der Wissenschaften und der Literatur | Mainz, Germany

Launched in 2013, the project *The School of Salamanca. A Digital Collection of Sources and a Dictionary of its Juridical-political Language* is establishing a collection of more than a hundred sources from Iberian theologians and jurists of the 16th and 17th century. These texts deal with political and juridical topics and the collection of sources is supplemented by a dictionary that comprises, next to biographical information on the authors of the sources, the development of central terms and ideas of the Western history of political and legal ideas, as it is reflected in the source texts.

Both parts, the sources and the dictionary, will be published under Open Access conditions. In the beginning of 2016, the project’s web application will be launched online with the first batch of sources and dictionary entries as TEI-XML texts plus corresponding facsimile images. At the moment, we are running a proof-of-concept mechanism and some experiments, which at the time of release will be constituting a Linked Open Data interface to our data along with a SPARQL-Endpoint.

At the beginning of the presentation the project’s rationale and web application will be introduced shortly. But the focus of the presentation will be on giving insights in the workflow, the decision making process and the implementation of the LOD mechanisms, that have been realized:

1. The first aspect accentuated in the presentation is the modelling of the information contained in our TEI-XML data within a Linked-Data-environment: Which TEI elements or attributes are assigned to which objects and predicates of which ontology? How are these assignments processed in order to offer the data as semantic data? Problems that we are dealing with are the questions of how to record the temporal dimension of much biographical information (see Ramos, 2009; Mynarz, 2013), and how to cope with alternative values like e.g. conflicting data about the date of birth of a person? Do considerations such as these affect our main objectives, e.g. the TEI-scheme or the collection of data?

Here is what has been settled up to now: We will offer semantic data about the sources and about the authors of the collection. The ontologies we use are mainly the foaf-, bio-, relationship-, and SPAR-ontologies (see, among others, Peroni, 2014). The TEI-data will be transformed into RDF by the xtriples webservice (Schrade, 2015).

2. The second highlighted aspect concerns the very networking of the data itself and its utilization in the project's infrastructure. This concerns technical issues, such as the questions of which services and resources should be – directly or indirectly – provided in order to offer our data for external reuse? It concerns issues of academic strategies such as negotiations with favored partner projects and data providers over the data they expose and the interfaces they provide. But it also concerns scholarly questions, such as eventual opportunities to handle new research questions, or to handle questions in new ways, opened up by the integration of our data with the data of other LOD-providers. In which form could or should such expanded possibilities be provided on the publicly accessible website? What stance are we taking on rights management and quality insurance of federated queries/data?

Again, here is the current status: We have mechanisms for the resolution of resource URIs, for content negotiation and for dumping the complete triple store as well as a SPARQL endpoint in place. We are elaborating federated queries responding to specific, concrete research questions. And we are still in the process of evaluating the Linked Data Fragments (Verborgh, 2014) mechanism. We are in contact with several projects the data of which would nicely complement our own (Schmutz, 2008; Sytsma, 2010; Bullón, 2012; Mrozik, 2016). And we are probing possibilities of offering a configurable interface to (federated) querying to our users and of rendering network information visually in our web application. Except most likely for the last point (due to time constraints), the talk will present the state of affairs we will have achieved in summer 2016.

The presentation will thus point out conceptions and their implementations of linking TEI resources into the semantic web, difficulties encountered and needs still left open by scholarly research questions.

Bibliography

Bullón, Xavier Agenjo (2012). Introducción: la Biblioteca Virtual de la Escuela de Salamanca y Linked Open Data. http://dx.doi.org/10.18558/FIL (accessed 8 April 2016).

Ciotti, Fabio et al. (2014). TEI, Ontologies, Linked Open Data: Geolat and Beyond. https://jtei.revues.org/1365 (accessed 8 April 2016).

Eide, Øyvind (2015). Ontologies, Data Modeling, and TEI. https://jtei.revues.org/1191 (accessed 8 April 2016).

Meroño-Peñuela, Albert et al. (2014). Semantic Technologies for Historical Research. A Survey. http://www.semantic-web-journal.net/system/files/swj588_0.pdf (accessed 8 April 2016).

Mrozik, Dagmar (2016). The Jesuit Science Network (Wuppertal/Berlin, being established in 2016). http://jesuitscience.net/ (accessed 8 April 2016).

Mynarz, Jindřich (2013). Capturing temporal dimension of linked data. http://blog.mynarz.net/2013/07/capturing-temporal-dimension-of-linked.html (accessed 8 April 2016).

Pattuelli, Cristina M. (2012). FOAF in the Archive: Linking Networks of Information with Networks of People. Final Report to OCLC. http://www.oclc.org/content/dam/research/grants/reports/2012/pattuelli2012.pdf (accessed 8 April 2016).

Peroni, Silvio (2014). The Semantic Publishing and Referencing Ontologies: 10.1007/978-3-319-04777-5_5.

Ramos, Michele R. (2009). Biography Light Ontology. An Open Vocabulary For Encoding Biographic Texts. http://metadata.berkeley.edu/BiographyLightOntology.pdf (accessed 8 April 2016).

Romanello, Matteo and Pasin, Michele (2013). HuCit. https://bitbucket.org/56k/hucit/ (accessed 8 April 2016).

Ruiz-Iniesta, Almudena and Corcho, Oscar (2014). A review of ontologies for describing scholarly and scientific documents. http://ceur-ws.org/Vol-1155/paper-07.pdf (accessed 8 April 2016).

Schmutz, Jacob (2008). Scholasticon. Ressources en ligne pour l'étude de la scolastique moderne (1500-1800): auteurs, sources, institutions. http://scholasticon.ish-lyon.cnrs.fr/Presentation/index_fr.php (accessed 8 April 2016).

Sytsma, David (2010). Post-Reformation Digital Library. http://www.prdl.org/ (accessed 8 April 2016).

Verborgh, Ruben et al. (2014). Querying Datasets on the Web with High Availability. In: Proceedings of the 13th International Semantic Web Conference: 10.1007/978-3-319-11964-9_12.

Wettlaufer, Jörg et al. (2015). Semantic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science: 10.1093/llc/fqv047.

Schrade, Torsten (2015). xTriples. A generic webservice to extract RDF statements from XML resources. http://xtriples.spatialhumanities.de/index.html (accessed 8 April 2016).

Preserving Ireland's Digital Cultural Identity towards 2116

Sharon Webb
sharon.webb@sussex.ac.uk
Sussex Humanities Lab, University of Sussex

Rebecca Grant
r.grant@ria.ie
Digital Repository of Ireland

1 Introduction

The Digital Repository of Ireland (DRI) is a national Trusted Digital Repository for Ireland's social and cultural data, accredited by the Data Seal of Approval. The repository links together and preserves historical and contemporary data held by Irish institutions, providing a central internet access point and interactive multimedia tools. In June 2015, the Digital Repository of Ireland was publicly launched at the *Digital Preservation for the Arts, Social Sciences and Humanities* (DPASSH) conference. At this conference, three organisations were presented with the DRI 'Decade of Centenaries Digital Preservation Award', the culmination of a six month project aiming to provide support and training to owners of digital collections relating to the Irish decade of centenaries commemorations. This paper will discuss the implementation of this Irish Research Council-funded project which allowed staff from the DRI to engage with collection owners and provide digital preservation and digitisation services and training. In this paper we outline the aims of the project, the methods by which we engaged with relevant collection owners, and how our findings have helped to determine the status of digital preservation in Irish heritage organisations.

2 Background

In June 2015, the first DPASSH conference was hosted in Dublin by the DRI, to highlight the technical, cultural, and social problems, challenges, and opportunities of long-term digital preservation in the arts, social sciences and humanities. The conference theme, 'Shaping our Legacy: Safeguarding the Social and Cultural Record', was developed to reflect concerns within the community that digital cultural heritage is at risk of destruction. The conference's call described the destruction of the Irish Public Records Office (IPRO) in 1922 during the Battle of Dublin. As the Irish Civil War broke out, a priceless archive containing a thousand years of Irish history was destroyed.

The destruction of the IPRO was used as an analogy to highlight the vulnerability and fragility our digital cultural heritage in the long-term. The challenge, however, is that "long-term", in this context, comes nowhere near the thousand years of history housed in the purpose-built store of the IPRO; as Jeff Rothenberg (1999) states, 'digital information lasts forever—or five years, whichever comes first.'[1]

It is within this context that we sought expressions of interest from custodians of heritage material relating to the Decade of Centenaries (DC) who wished to digitally preserve their holdings. Funded by the Irish Research Council's New Foundations Award, we aimed to assess the scale of vulnerable digitised collections related to the DC in Ireland, to provide support in digitally preserving those collections, and to create a centralised access point to encourage their wide dissemination. Importantly, we wanted to raise awareness of the issues related to digital preservation and provide resources, best practice advice and guidance to all applicants. We also planned to have an impact on the community beyond the span of the funded project, providing equipment and training to continue to support the digital preservation of Irish cultural heritage.

3 Methodology

Our call for expressions of interest was announced in December 2014 and was circulated by the DRI community. The call sought collections that were partially or fully digitised and described, and which contributed to the national narrative on the period 1912-22. The winners were offered resources to ensure the digital preservation of their collections, including staff time from professional archivists and librarians, digitisation services, metadata creation, and the ingestion of content into the DRI for long term preservation.

Interested collection owners were invited to submit a detailed application form, providing information on the types of digital assets in their collection, the volume, current storage provisions, and its connection to the DC.

Eight proposals were received, and through the collection assessment and selection phase, three were chosen: the Irish Capuchin Provincial Archives (The Capuchins and the Irish Revolution),[1] the Dublin City Archives (Dublin City Electoral Lists, 1915),[2] and the National Irish Visual Arts Library (Michael Healy Collection).[3]

Our assessment procedure enhanced our understanding of the types of relevant digital assets held nationally, and gave us insight into the challenges faced by collection owners in ensuring that their content is digitally preserved.

The second phase of implementation included a scoping exercise and requirements analysis for each collection, the creation of a project plan and allocation of resources. Requirements gathering was conducted through interviews with collection owners, undertaken by the DRI's Requirements Manager and a Digital Archivist or Librarian in all cases. Work plans were created with tasks including digitisation, metadata creation and standardisation, ingest preparation and collection creation, review and publication. We worked closely with each collection owner to ensure that the processes and workflows we created could be repeated, and that these were reflected in our existing guidelines.

Following the completion of the work plans, a digital preservation workshop was held at the Royal Irish Academy to ensure that collection owners were trained in the procedures required to prepare their content for ingestion into the DRI Repository - all applicants to the original call were invited.

4 Findings

From the initial submissions to the call for expressions of interest, it was clear that there is a need in the community, not only for preservation services, support and advice, but also for digitisation support. Digitisation services were requested by nearly all of the applicants, and in some cases digital preservation was not included in the application. It appears that Irish archives lack resources (i.e. staff time and equipment) for digitising their collections. While perhaps not surprising, the technical infrastructure required to provide robust digital preservation was also not available to any of the participants. This confirms the need for shared infrastructures, or indeed shared strategies on preservation, on a national level.

Regarding the three selected collections, it was found that while metadata standards had been applied, they were in some cases customised according to the needs of the collection owner. Furthermore, although ISAD(G) compliant descriptions had been used by two collection owners, these could not be exported from the cataloguing software as EAD-XML, and needed to be manually marked up using an XML editor. These limitations create a barrier to metadata exchange and indicate that some archives are not planning for interoperability with other collections and/or repositories. While this was indicated in our 2012 report, Digital Archiving in Ireland, this project highlighted the practical difficulties involved in overcoming these barriers to interoperability.[2]

The DRI guidelines and workflows were tested by the process of preparing content for preservation in the Repository and found to be comprehensive and robust. However, the underlying knowledge in the community regarding standardised metadata creation and the principles of digital preservation was not well developed. The preservation workshop held at the RIA was booked to capacity with a waiting list, and training in basic digital preservation and metadata preparation was identified as a requirement for our stakeholder community.

5 Conclusion

The award has allowed the DRI to engage with a number of new stakeholder organisations who had not previously undertaken any digital preservation processes for their collections. Through the award, and the subsequent preservation workshop, the team worked with seven organisations who had not previously deposited content with the Repository. As well as providing training to allow participants to deposit with DRI, advice was also provided on smaller scale, in-house preservation practice which participants could bring back to their organisations.

The DRI believe that digital heritage is at risk of destruction and loss if action is not taken. Digital decay is the gradual decay of digital content. The solution is digital preservation – active ongoing data management. Long-term digital preservation is concerned with providing sustained access to digital objects and content and requires that institutions are cognisant of the processes and procedures required to ensure the form, as well as functionality, of digital objects.

DRI actively engaged with our designated community and have sought participation throughout our phases of development - from our requirements analysis phase, through to user acceptance testing and content population. Therefore, while the DC call provided an opportunity to preserve and publish content it also provided a platform from which to engage with an important national programme of events, communicate with a wider audience and test our user guidelines and documentation. Crucially, it also emphasised the fact that long-term preservation is a socio-technical problem - the solution requires not just digital infrastructures but advocacy, industry and societal engagement, and cooperation with content owners.

In addition to the practical aspect of this work we wanted to highlight the need for the national programme of commemorations to be cognisant that digital collections or projects created now, should be considered as historical artefacts for future historians. That is, our current digital commemorations should be preserved for future use and analysis. Commemorative events (both online and offline) are performative acts of nationalism and are an integral part of how we understand both our current and future selves. Current national projects should consider how they are preserving Ireland's digital cultural identity for 2116.

Acknowledgements

We acknowledge the support of the Irish Research Council's New Foundations Programme.

Bibliography

Rothenberg, Jeff (1999). Ensuring the longevity of Digital Information. http://www.clir.org/pubs/archives/ensuring.pdf

O'Carroll, Aileen and Webb, Sharon (2012). Digital Archiving in Ireland: National Survey of the Humanities and Social Sciences. Maynooth University. DOI: 10.3318/DRI.2012.1

Notes

[1] http://dx.doi.org/ 10.7486/DRI.95944s31k

[2] http://dx.doi.org/ 10.7486/DRI.9593zg12h

[3] http://dx.doi.org/ 10.7486/DRI.95944s32v

Faceting and Mining Network Graphs

Claude Willan
cwillan@princeton.edu
Princeton, United States of America

The network graph is one of the best-known and overdetermined of all data visualizations. And it suffers, more than most such modes, from the problem of fetishization. When non-specialists see network graphs, which are information-heavy, aesthetically appealing, cognitively suggestive, and yet curiously hard to read, their first reactions are often along the lines of "Oh wow!", "Cool!", or "Neat!". The corollary to admiration, for members of the academy, however, is often distrust.

The problem is two-fold. Critical thinkers are justly skeptical of enthusiasm, since enthusiasm can foreclose critical engagement. But the network graph's prominence as a mode of data visualization, as featured in new outlets like the New York Times, or in prominent projects like Mapping the Republic of Letters or Six Degrees of Francis Bacon , makes it uniquely vulnerable to the hermeneutic of suspicion (see Elijah Meeks, "The Digital Humanities as Lightning Rod"). If the digital humanities are still distrusted in some more conservative corners of the academy — such as mine, eighteenth century english literary manuscript studies — then network graphs, because they have become synecdochic for digital humanities writ large, are doubly distrusted.

This short paper addresses three states of addressing this skepticism head-on, embedding network graph literacy in the context of a larger disciplinary argument. Those three stages correspond to three key perspectives practitioners of network analysis can assume when wanting to persuade skeptics of the probative value of network graphs, to show how condign those graphs are to traditional "analogue" analysis, and to build a persuasive argument within your own field.

Those three perspectives are: targeted data gathering; graph design, and argument design.

Much of my own work centers on networks of distribution of Jacobite manuscript poetry from 1688 — 1750. Once I knew that I would be using network graphs to illustrate the richness of the social and material facts I had uncovered, I started enriching my dataset with an eye to encoding metadata to reveal those richnesses as precisely as possible. Those metadata categories include size of manuscript, number of copies of each poem in circulation, names of manuscript collectors, languages contained in each manuscript, dates of manuscript, and so on. I entered this metadata on each future node in my two sets of network graphs (one of individual poems, one of manuscripts).

Graph design, the next stage, was just as important. Having settled on a layout in Gephi that I felt reliably showed relations among objects in a way that I knew I could explain simply (such as proximity as a function of similarity; distance as an index of relative dis-similarity) I then worked to combat the visual fetish character of the network graph by partitioning the nodes multiple times while keeping them within the same layout. This meant that one my audience had had time to accustom themselves to the layout, successive graphs with differing color schemes, node sizes, and edge weights would show them the facets, the richness, of the dataset. By showing the same graph layout faceted in multiple different ways, my audience could comprehend the layout as a function of multiple and overlapping functions.

Argument design, the last stage, is a way of making the procession of network graphs an intuitive progression of visual arguments that draws the audience into, and educates them about, the rhetoric of the network graph precisely by showing its malleability and, in some ways, its very partiality. By showing the protean capacities of the network graph, by showing the extent to which each network graph is itself a reading, skeptics more at home with paleography and stemmatology are able to grasp quite intuitively the resource that is faceted network graph: a flexible and powerful tool to show the richness and interrelatedness of large datasets based on archival discoveries very like their own.

This short paper will include slides of my own faceted network graphs to show how we can use this form of argumentation as a kind of nested pedagogy.

I enclose here a collection of faceted and mined network graphs made for a colleague's project on Modernist journals. These bimodal graphs show the interrelation of different contributors to various roughly contemporary journals. By showing the same layout several times, but using node color, node size, edge weight and color, we are able to see:

- which nodes are journalists and which are journals;
- nodes clustered with a granularity of 1.0 (fewest communities of the largest size);
- nodes clustered with a granularity of 0.23 (many smaller communities);
- which journalists contributed the most (edge weight), which journals had the most contributors (node redness and size), and
- which journalists contributed to the most journals (by node size).

These five network graphs taken together not only supply a much richer picture of the system of writing and publication than a single graph would; by showing the range of information that can be displayed on that graph with a series of simple adjustments, the reader is shown how the graph is a product not of vulgar scientism but of intentional humanism.

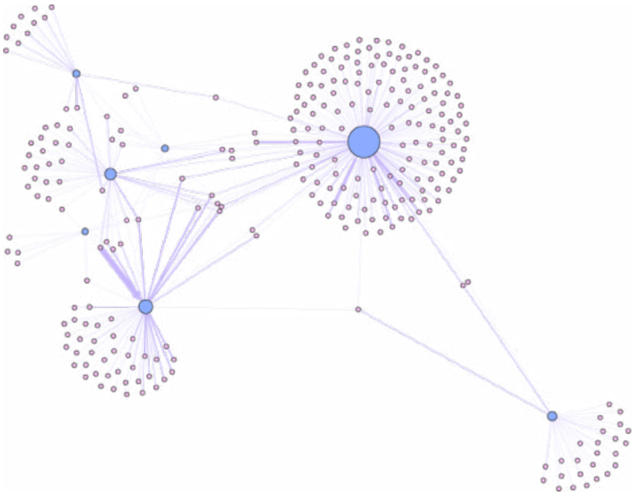


Illustration 1: Lilac nodes are journalists; blue are journals

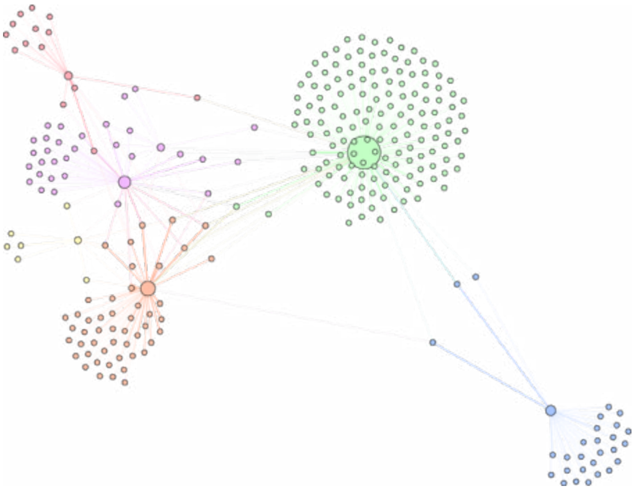


Illustration 2: Modularity at granularity 1.0; the fewest communities of the largest size.

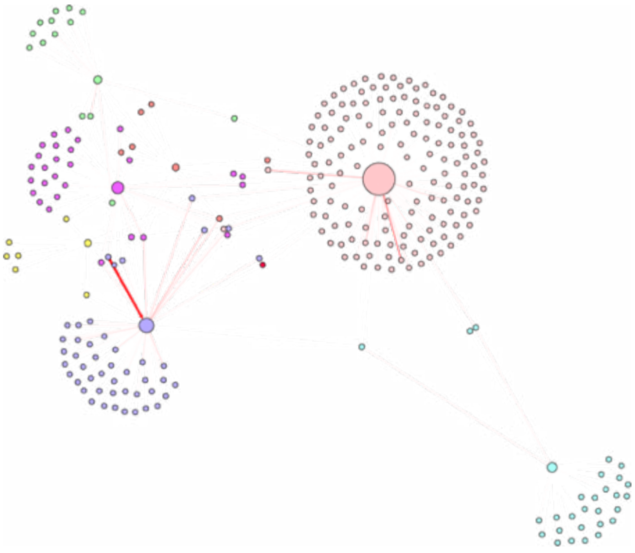


Illustration 3: Modularity class 0.23; more communities of a smaller size.

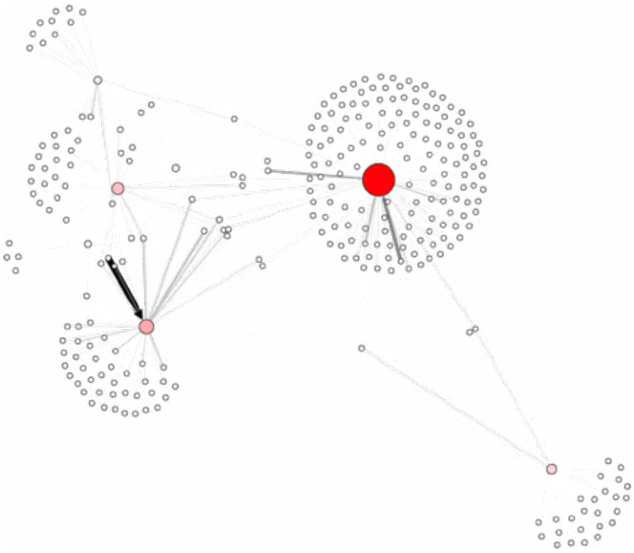


Illustration 4: Edge weight shows number of of contributions by a journalist; node redness and size indicates number of contributors.

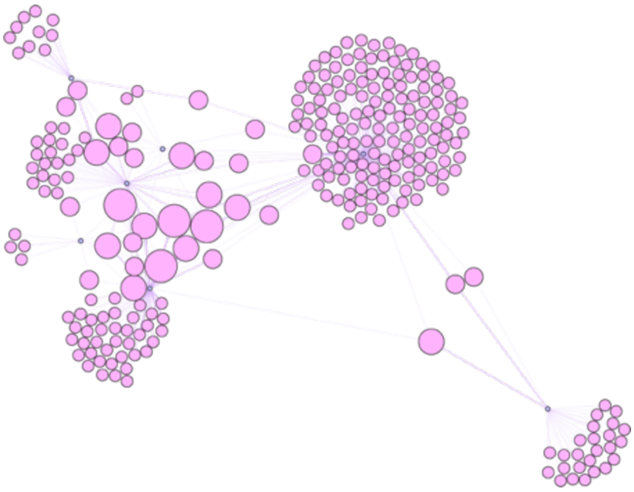


Illustration 5: Node size indicates number of journals to which a journalist contributed.

Bibliography

Mapping the Republic of Letters. http://republicofletters.stanford.edu/ (accessed March 7th 2016).

Meeks, E. (2012). The Digital Humanities as Lightning Rod. https://dhs.stanford.edu/the-digital-humanities-as/digital-humanities-as-lightning-rod/ (accessed March 7th 2016).

Six Degrees of Francis Bacon. http://sixdegreesoffrancisbacon.com (accessed March 7th 2016).

Tackling Terms In Furetiere's 'Dictionnaire Universel'

Geoffrey Clive Williams
williams@univ-ubs.fr
Université Grenoble Alpes, France

First published posthumously in 1690, Furetière's *Dictionnaire Universel* had aroused controversy well before its publication. Nevertheless, it was quickly followed by an enlarged and corrected version edited by the protestant scholar Henri Basnage de Beauval. It is this extended version of 1701, with its broad coverage of terminological language that is our subject.

The Furetière project seeks to render the entire 1701 dictionary available as an open access digital resource in an XML-TEI compliant format. Given the size of the task, and the current total lack of finding with little hope in the short term, the first stage will be an attempt to map and describe terminological coverage by reference to a small number of themes, namely architectural, legal and maritime terminology. This paper will demonstrate the mapping procedure being carried out using the Atlas Ti Computer-Assisted Qualitative Data Analysis Software (CAQDAS) and the building of a model for the TEI encoding.

Historical background

Furetière was both a member of the French Academy and participant in the dictionary building team, hence the uproar at his publishing what was seen as a rival dictionary. This explains why a Catholic priest should end up being published by the protestant publisher in the Netherlands, Arnaud Leers. The dictionary was a success, but still needed much revising. This was carried out by the French émigré and scholar, Henri Basnage de Beauval. Despite being the most complete edition, only one attempt at digitising was ever made (Wionet and Tutin, 2001). Our aim is to encode the Basnage dictionary but with cross references to the edition of 1690 and the publication of Corneille (1694) so as to see how the terminological entries evolved and to what extent complementary information can be found in the Corneille dictionary (Williams, Forthcoming).

Using CAQDAS

One important task for the Digital Humanities community will be to bring or adapt existing tools to disciplines that are often less digitally aware. Thus, in using a CAQDAS tool to explore the three dictionaries, we aim not only to use a very powerful tool to allow pre-digitising analysis and mapping of data, but also to bring this technology into the sphere of literary analysis.

CAQDAS were essentially created to meet the needs of sociologists carrying out rigorous qualitative analyses on data from multiple supports. Literary specialists also carry out qualitative analyses, but often tend to use highlighters to work on printouts of PDFs. By using a CAQDAS to work through a document as big and as complex as the Furetière's dictionary, we aim to show how an electronic highlighter that allows coding and network analyses can be used in humanities research, and particularly in our own.

There are a number of commercial CAQDAS on the market, and only one open-source tool. Although open-source tools are important to the community, we have adopted Atlas Ti as being a very powerful tool that is evolving rapidly with new functions.

The 1701 Basnage edition is over 4000 pages of very tight text. Text quality is poor which precludes use of OCR. Whilst it is structured, we are in a period of great experimentation in dictionary compilation leading to a complex meta and microstructure. The only way forward is to read the text. Using Atlas Ti, we skim the pages so as to locate and code entries designated as terms, to find the different introductory formulae, as well as spotting potential search formulae for unmarked terms, and to list the domains and crafts to which they belong. The coding system allows us to carry out a bottom-up conceptualisation of the dictionary with the quotations allowing us to create a headword list that links directly to the entries in the PDF file. Knowing the terminological domains covered by the dictionary is interesting in itself, but it also gives further keys for reaching unmarked terms. Networks allow us to reorganise domains and crafts into groups, such as legal terminology – represented by several domain names-and maritime terminology, which is often closely linked to language of fortification, architecture and law. Atlas Ti allow to output data in a machine-readable format so that the headword list could be transformed as an organised lexicon for use in the XML encoding process.

XML TEI

Use of Atlas Ti as an electronic highlighter with all its coding and management functions does permit a full qualitative analysis of the data. However, it is still not possible to share the data itself. Digitising the entire dictionary is a mammoth task that is only feasible using crowd sourcing over a long period of time. Marking up only thematically designated terminological domains allows us to create and test a model for the data as well as making available machine readable data rapidly. The dictionary is unique in having definitions accompanied by lengthy examples and citations thus providing both technical and phraseological information that will only be retrievable using in-depth mark-up. Whereas Wionet and Tutin (op.cit) marked up one letter, our plan is to attempt to follow the terminology through the entire dictionary.

The first stage consists of marking up terms from a small number of highly productive thematic fields isolated using the CAQDAS analysis. The first field to be explored in depth is related to maritime activities as designated by 'terme de marine' (maritime term) so as to open a collaboration with French historians so as to compare an élite vision of naval terminology as compared by the situation in a major naval port, that of Lorient which founded in 1666 by the Compagnie des Indes, and then became a major naval base from 1703.

Our aim is to illustrate the decisions taken and how these affect output through visualisation, but also analysis using linguistic analysis tools as TXM and a database system as BaseX. Once more advanced, the data will be put online using a query interface. At the moment, we are sharing code on GitHub under the name Basnage.

Mark-up is being carried out using Oxygen so as to mark-up using the TEI guidelines and use XQuery to ensure consistency. Entries can be extremely complex, as will be illustrated by reference to the verb *Abatre*. This means that the TEI guidelines do not always adapt easily to what is found in the text. However, we are endeavouring not to customise the guidelines so as to retain full compatibility with other dictionaries and allow easier linking with source texts.

The dictionary tends to group words orthographically so that there is a main headword in large capitals, which also carried the grammatical information, but then a series of subentries that generally have the headword in small capitals. These subentries may include the specialised terminological usage. To complicate affairs, polysemy is illustrated within an entry with short comments and examples. To handle this, we are using <superEntry> to group the whole, and then <entry> for what might be considered as subentries. <Sense> is used to cover polysemy within a given entry. The main entry for ABATRE has three main senses illustrated by a series of synonyms, with each sense accompanied by numerous examples, and occasionally a citation with bibliographic reference. The examples frequently contain collocations that activate a particular synonym, it would be useful to mark-up and illustrate this. Similarly, citations generally only give a link to a person, often via an abbreviated name, as in MEN for Ménage or ABL for M. d'Ablancourt. These are generally listed at the beginning of the dictionary, but it is not always the case and sometimes the abbreviations are inconsistent. Sometimes, even if a text is not named, it is possible to link to a source document. Given that Basnage was a prolific letter writer at the centre of network of European scholars, and the author of the *Histoire des Ouvrages des Savants*, and important task will be to cross link to this valuable source of information.

Collocation mark-up means that we can link dictionary analysis to words in a wider context as fund in Frantext or, when a text has already been digitising by using a language analytical tool as TXM. Mark-up also aims to make an onomosiological analysis possible using BaseX rather than simply presenting the data in a linear semasiological format.

Conclusion

This is a mammoth mark-up task which we believe is rendered easier by mixing tools so as to permit on-going analysis whilst gradually digitising the whole into XML compliant TEI. This strategy means that other scholars can use data without having to wait for the entire digitalisation process to be completed. In so doing, we seek to explore data whilst collaborating in the dissemination and improvement of digital tools and also a contribution to the art of digital mark-up of early dictionaries.

Bibliography

Williams, G. (Forthcoming). Le temps des termes: les termes et la phraséologie dans les dictionnaires du 17 siècle. In De Giovanni, C. (Ed.), *Fraseologia E Paremilogia: Passato, Presente E Futuro*. FrancoAngeli: Milan.

Wionet, C., Tutin, A. (2001). *Pour informatiser le Dictionnaire universel de Basnage (1702) et de Trévoux (1704): Approche théorique et pratique*. Honoré Champion: Paris.

Digitale Tools und Methoden für die geisteswissenschaftliche Forschung praxisnah erklärt: Ein neues Format im Test.

Tanja Wissik
tanja.wissik@oeaw.ac.at
Austrian Academy of Sciences, Austria

Claudia Resch
claudia.resch@oeaw.ac.at
Austrian Academy of Sciences, Austria

Aufgrund der technischen Entwicklungen in den letzten Jahrzehnten sind GeisteswissenschaftlerInnen in ihrem Forschungsalltag vor neue Herausforderungen gestellt. Während es für den wissenschaftlichen Nachwuchs in Österreich eine Reihe von curricularen Angeboten sowie Summer Schools im Bereich der Digitalen Geisteswissenschaften gibt[1], sind erfahrene Forschende meist auf sich gestellt, wenn es darum geht, sich forschungsrelevante ICT-Kompetenzen für individuelle Fragestellungen anzueignen: Gemäß der von DARIAH[2] initiierten Umfrage (Schneider / Scholger in Druck) im Jahr 2014/2015 gaben mehr als 50 Prozent der befragten Forschenden in Österreich an, dass Weiterbildungsangebote zum Thema DH-Methoden und Werkzeuge und wie diese ihre eigene Forschung verbessern könnten und tendenziell wichtig oder sehr wichtig für ihre Arbeit wären.

Das Austrian Centre of Digital Humanities (ACDH) an der Österreichischen Akademie der Wissenschaften sieht es als eine seiner Aufgaben, DH-Inhalte und -kompetenzen zu vermitteln und die Forschenden dabei zu unterstützen, das Potenzial digitaler Methoden und Werkzeuge für ihre konkreten Forschungsprojekte zu nützen. Aus diesem Grund werden am ACDH unterschiedliche Formate der Wissensweitergabe erprobt und evaluiert. In unserem Kurzvortrag stellen wir eines dieser Formate, die ACDH Tool Gallery[3], als Fallstudie vor und berichten über deren Konzeption und Etablierung als außeruniversitäres Weiterbildungsangebot.

Anders als in herkömmlichen Vortragsreihen, in denen praktischen Aspekten weniger Bedeutung zugemessen wird, stellt die ACDH Tool Gallery das Experimentieren mit eigenen Daten in den Mittelpunkt. Das Konzept der ACDH Tool Gallery besteht darin, zwei Gruppen von Expertinnen und Experten zusammenzubringen: einerseits jene, die sich mit der Entwicklung von Tools beschäftigen und diese bereitstellen, und andererseits jene, die in ihrer geisteswissenschaftlichen Fachdisziplin ausgewiesen sind, und diese Tools in Zukunft verwenden möchten. Bei der Auswahl der Kurzreferate am Vormittag wird diesem Konzept insofern Rechnung getragen, als dass sowohl IT-ExpertInnen als auch GeisteswissenschaftlerInnen, die diese Tools bereits für ihre Forschung einsetzen, als Vortragende eingeladen werden. Die eintägigen Veranstaltungen sind jeweils einem Themenkomplex gewidmet: Die Tools, die am Vormittag in Einführungsvorträgen und Präsentationen vorgestellt werden, können nachmittags von den TeilnehmerInnen erprobt werden, indem diese Schritt für Schritt von der Installation bis zur Anwendung begleitet werden. Idealerweise hat jeder der TeilnehmerInnen ein eigenes, individuelles Set an Daten am Laptop vorbereitet, an dem die Tools getestet werden. Mit der Teilnahme an der Veranstaltung kann eine Kursbestätigung erworben werden.

Die ACDH Tool Gallery versteht sich als Angebot für EinsteigerInnen im Bereich der Digital Humanities und vermittelt einen Überblick sowohl zu bewährten als auch zu neuesten Tools, die zur Verfügung gestellt werden. Im Vordergrund steht der Austausch zwischen AnbieterInnen und potentiellen AnwenderInnen: Gemeinsam kann eine Einschätzung darüber erfolgen, ob und inwieweit ein jeweiliges Tool zur Beantwortung individueller Forschungsanliegen geeignet ist. Ausreichend Zeit zur (spontanen) Diskussion ist die Voraussetzung zum Gelingen dieses innovativen Formats.

Die Zielgruppe sind ForscherInnen aller geisteswissenschaftlichen Disziplinen sowie zum Teil MitarbeiterInnen von wissenschaftlichen Einrichtungen wie Bibliotheken und Archiven. Die ACDH Tool Gallery ist ein unentgeltliches Angebot, das sich nicht nur an ForscherInnen der Österreichischen Akademie der Wissenschaften richtet, sondern allen Interessierten offensteht. Damit dieses Format gelingt, braucht es intensive Vorbereitung auf allen Seiten: 1. liegt es an den TeilnehmerInnen zu überlegen, an welchen Daten sie das jeweilige Tool zu testen beabsichtigen; 2. leisten die Tool ExpertInnenen Vorarbeit, indem sie einen Zugang zu ihren Tools schaffen; und 3. gibt das ACDH die konzeptionellen und organisatorischen Rahmenbedingungen für dieses Format vor. Da es sich um ein neues Angebot handelt, wird die ACDH Tool Gallery intensiv über die ACDH-Website, die dha-Website, die ÖAW-Website, diverse Mailinglisten und Social Media (Facebook and Twitter) beworben.

Die ACDH Tool Gallery wird dreimal im Jahr angeboten und hat bislang bereits zu den Themen "Automated Recognition and Transcription of Handwritten Documents", "Basic Text Enrichment - TreeTagger for DH-Application" und "Semantic Web Tools" stattgefunden. In einem Bewertungsbogen wurden die TeilnehmerInnen jeweils nach der Veranstaltung befragt, wie sie dieses neue Format einschätzen, mit welchem Vorwissen sie daran teilgenommen haben, was sich verbessern ließe und welche weiteren Tools relevant für ihre Forschungen wären, um künftig im Rahmen der Tool Gallery vorgestellt zu werden. Im Fragebogen wurden außerdem anonyme personenbezogene Daten (z.B. Alter, Geschlecht, Beruf, Herkunftsinstitution) erfasst.

Die Ergebnisse dieser Fallstudie werden im Kurzvortrag präsentiert. Vor dem Hintergrund dieser Erfahrungen möchten die Autorinnen diskutieren, was daraus abgeleitet werden kann und Empfehlungen geben, wie ein attraktives, außeruniversitäres Weiterbildungsangebot, das den disziplinären Anforderungen[4] unterschiedlicher Forschenden gerecht wird, gestaltet werden könnte.

Bibliographie

Bauer, B., Ferus, A., Gorraiz, J., Gründhammer, V., Gumpenberger, Ch., Maly, N., Mühlegger, Johannes M. Preza, J. L. Sánchez Solís, B., Schmidt, N., Steineder, Ch. (2015). *Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung – Report 2015*. Version 1.2, DOI: 10.5281/zenodo.32043. Online auch unter: http://phaidra.univie.ac.at/o:407513 (letzter Zugriff 29.10.2015).

Sahle, P. (2013). *DH Studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities*. DARIAH-DE Working Papers Nr. 1. Göttingen: DARIAH-DE. URN: urn:nbn:de:gbv:7-dariah-2013-1-5, Anhang BA- und MA-Studiengänge, pp. 30-31 (letzter Zugriff 29.10.2015).

Schneider, G., Scholger, W. (2015). In: Dallas and N. Chatzidiakou (Eds.), *DARIAH survey on scholarly practices and needs of European humanities researchers in the digital environment 2014-15*, Technical report, Athens: Digital Curation Unit. Forthcoming, in Druck: Austria.

Notes

[1] Vgl. Liste von DH-Studiengängen im deutschsprachigen Raum verfügbar unter http://www.cceh.uni-koeln.de/Dokumente/BroschuereWeb.pdf [letzter Zugriff 29.10.2015]; Sahle, Patrick (2013): "DH Studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities". DARIAH-DE Working Papers Nr. 1. Göttingen: DARIAH-DE. URN: urn:nbn:de:gbv:7-dariah-2013-1-5, Anhang BA- und MA-Studiengänge, 30-31 [letzter Zugriff 29.10.2015]; Dariah.eu Digital Humanities Course Registryhttps://dh-registry.de.dariah.eu/ [letzter Zugriff 29.10.2015]; Tabellarischer Vergleich der Studiengänge. In: Thaller, Manfred (Eds.), (2015). Digital Humanities als Beruf. Fortschritte auf dem Weg zu einem Curriculum. Akten der Dhd Arbeitsgruppe „Referenzcurriculum Digital Humanities“ vorgelegt auf der Jahrestagung 2015, Graz 24.-27. Februar 2015, 123-125.

[2] Digital Research Infrastructure for the Arts and Humanities, http://dariah.eu/ [letzter Zugriff 29.10.2015]

[3] https://acdh.oeaw.ac.at/acdh/en/acdh-tool-gallery [letzter Zugriff 29.10.2015]

[4] Vgl. die Empfehlungen zur Implementierung eines österreichweiten Schulungsprogrammes zum Thema Forschungsdaten speziell im Bereich der Geisteswissenschaften bei Bauer, Bruno; Ferus, Andreas; Gorraiz, Juan; Gründhammer, Veronika; Gumpenberger, Christian; Maly, Nikolaus; Mühlegger, Johannes Michael; Preza, José Luis; Sánchez Solís, Barbara; Schmidt, Nora; Steineder, Christian (2015): Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung – Report 2015. Version 1.2, S. 70. DOI: 10.5281/zenodo.32043. Online auch unter: http://phaidra.univie.ac.at/o:407513

The Formation of Australia's Economic History Community, 1950–1970: A Multidimensional Network Analysis

Claire Wright
clairew@uow.edu.au
University of Wollongong, Australia

This paper adopts a social-deterministic view of intellectual history in order to link changes in social and professional interactions to the intellectual character of a community. The formation of Australia's economic history community between 1950 and 1970 is used to illuminate the link between these social and intellectual forces. The economic history field globally has experienced a number of changes over the twentieth century, and although some of these aspects are understood for the larger communities in the US and the UK (see Hudson, 2001; Lyons et al., 2008 for recent examples), Australia's economic history field remains neglected. Social connections within the community were driven by contextual factors such as the post-WWII expansion of higher education, the growing emphasis on research output and the development of domestic PhD programs. These connections then developed in a multi-dimensional way throughout the 1950s and 1960s, with co-location and collaborative relationships emerging simultaneously between individuals. It is found that the increased density and multiplexity of ties in the community contributed to the development of a distinctive intellectual approach.

This project combines the quantitative analysis of social and intellectual relationships between members of the field with the qualitative analysis of published works of economic history in this period. Though there is a long tradition of emphasising the importance of social interactions for the development of ideas (Kuhn, [1962] 1970; Mulkay et al., 1975; Whitley, 1984), there have been only limited attempts to visualise social relationships in intellectual history (for some examples, see Harvard University, 2015; Shakeosphere, 2015; Six Degrees of Francis Bacon, 2015), and even fewer attempts to relate changes in the social structure to changes in the intellectual character of the group.

Methodology

This project links the social and knowledge networks for this community. The social network has been visualised through co-location and collaboration networks. The knowledge network has been visualised through citation analysis and is discussed further through the qualitative analysis of texts. The collaboration and citation analyses are based on a selection of key texts of Australian economic history for the period 1950 to 1970. Texts have been selected from prior wide reading of the subject, with further guidance from secondary analyses that focus on the literary aspects of the field. This has determined the key ideas, approaches and debates for this community, which has informed the selection of texts.

Based on these texts, collaboration for this community has been recorded in three ways: co-authorship, contributions to edited works, and sub-authorship (acknowledgments). These networks are bonded-tie and valued, based on the number of separate texts that each pair of individuals collaborated on. Collaboration in this context represents ongoing, two-way interaction and the trading of theoretical insights and ideas (Laudel, 2002; Wang et al., 2014). Co-location has been used to map geographic proximity between two individuals, assuming that if they worked within the same university or faculty, they were more likely to have contact than those who were geographically disparate (Jaffe et al., 1993; Ponds et al., 2007). This network is also bonded-tie and valued: if two people were both employed by the same university for 2 years, their relationship is given a score of 2, and so on.

From the raw collaboration and co-location data, a network of key actors in the community has been determined to allow the combination of the different social networks. All authors of key texts and key collaborative works are included. Otherwise actors are included if they were involved in more than one type of social interaction in this community, for example if an individual was appointed to one of the key universities in this period, as well as involved in the sub-authorship network. From this, each social network map has been weighted based on the strength of that particular relationship. Co-location has been weighted by 0.5, because although geographic proximity may induce interaction between scholars, there is no guarantee. Sub-authorship and contributions to edited works are weighted by 2, and co-authorship is weighted by 4.

The knowledge network has been partially analysed through citations, as these represent the ideas that are shared between different authors and the place of a text in the wider context of the field (Leydesdorff and Amsterdamska, 1990; Newman, 2010; Siler, 2013). The citation network for this project has been coded manually, as texts are generally non-digitised books and articles – features that make the use of the Social Sciences Citation Index (SSCI) inappropriate here. This network is directed, indicating the one-way transfer of ideas, and is valued for the number of times in each text that the citation is made.

Supplementing the quantitative analysis of social and intellectual relationships is a qualitative component. The key texts in this community, as identified above, have been analysed using Lloyd's (1995) framework, which classifies works of economic history based upon the author's assumptions about how the economy operates (ontology) and the author's methodology for gaining knowledge about the economy (epistemology). By analysing both qualitative and quantitative aspects, this methodology is able to link social and intellectual interactions with the ideas that emerge in the author's texts.

Preliminary results

The data has been visualised with *NetDraw*. Although the community had some mobility between different locations, the co-location map shows that interactions were primarily between those in the same city. This includes those appointed to the Australian National University (ANU) (located in Canberra, and visualised as the large cluster on the right in figure 1), the University of Melbourne and Monash University (both located in Melbourne, and visualised as the cluster at the bottom of figure 1), or the University of Sydney and the University of New South Wales (both located in Sydney, and visualised as the cluster on the left of figure 1).

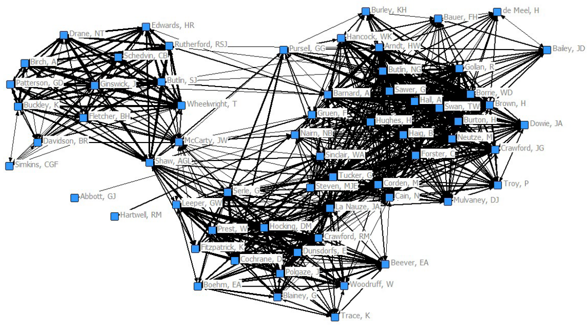


Figure 1: Co-location, 1950–1970

Collaboration was less geographically determined, though there was still greater interaction between those working at the same site. All co-authors were those who also had co-location ties (see figure 2). Contributors to edited works were also generally those that were geographically proximate to the editors in this period (see figure 3). Sub-authorship was the least geographically determined social network, responsible for most of the interactions between those in different cities (see figure 4). Having said that, the sub-authorship network was still clustered around the main institutions, with the ANU group experiencing the greatest frequency of sub-authorship in this period. Figure 5 shows that when each of these measures are combined, social interactions between members of the community in this period broadly followed co-location trends.

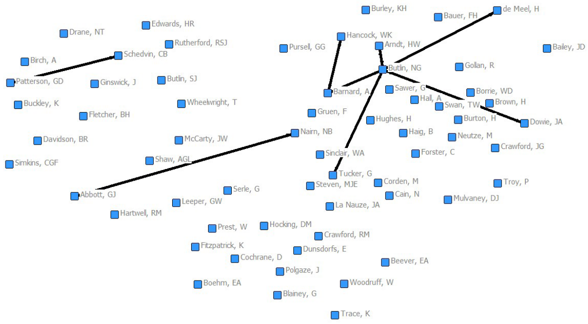


Figure 2: Co-authorship, 1950–1970

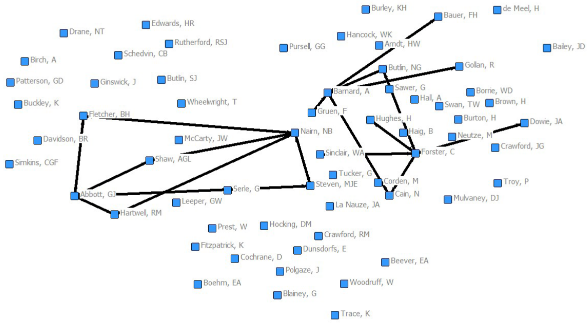


Figure 3: Contributors to edited works, 1950-1970

Figure 6 shows the citation network for this community, with authors of those texts included in the citation analysis highlighted. There was a broad tendency for those that had strong interactions in the social network to have similar citation patterns. A good example of this is NG Butlin, WA Sinclair, N Cain and AR Hall, located together towards the centre of figure 6. In the social network, these individuals were part of the large ANU cluster, with co-location and collaboration interactions throughout this period. There is evidence to support the counterfactual as well: that a lack of social interactions with the core of the community led to distinctive citation patterns. BR Davidson and E Dunsdorfs are good examples here, located on the fringes of figure 6 below, and also with limited interaction in the social networks above. However, this is a loose relationship, with non-social factors such as choice of research topic and approach to the subject also converging citation patterns.

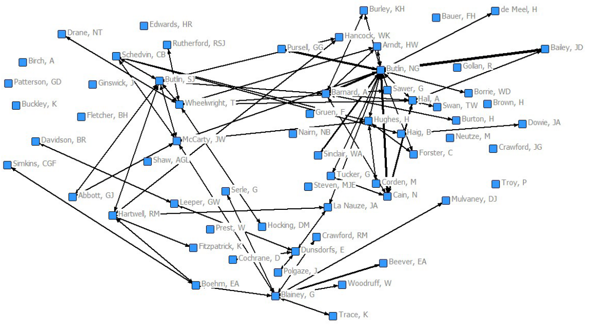


Figure 4: Sub-authorship, 1950–1970

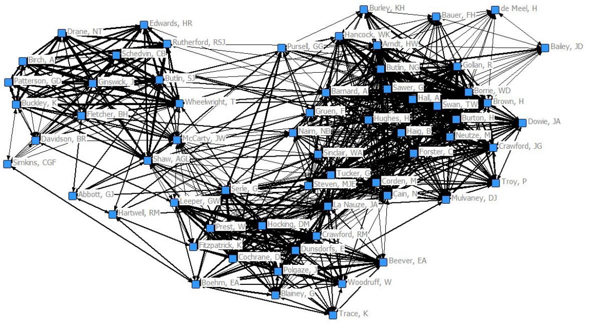


Figure 5: Multiple measures, 1950–1970

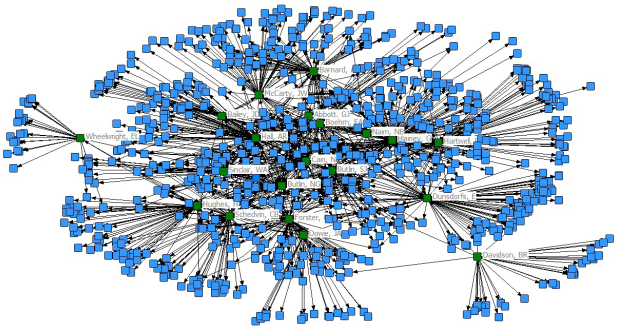


Figure 6: Citations, 1950-1970

The networks above show that there was a range of social and intellectual interactions in Australia's economic history field between 1950 and 1970, particularly structured around key institutions and prominent figures. This was accompanied by the development of a distinctive intellectual approach for the community. The orthodox approach to economic history in this period focussed on national income accounting, domestic determinants of growth, urbanisation, and the treatment of economic agents in an abstract and aggregate way. Core proponents of this approach had multidimensional interactions in the social network, mostly centred on NG Butlin and his colleagues at the ANU. A number of alternative approaches existed in smaller 'pockets' in the community, which were also structured by social interactions.

By combining the quantitative analysis of various social and intellectual relationships with the qualitative analysis of texts and ideas, this project links the social network and the knowledge network for Australia's economic history community. Not only does this provide a pioneering attempt to combine SNA with the more traditional methods for intellectual history, it analyses the Australian economic history field in a dynamic and multi-dimensional way. Key results suggest an association between social interactions and the citation patterns adopted by members of the community in this period. Key social and intellectual relationships between members of the field also led to the development of a distinctive intellectual approach for this interdisciplinary field.

Bibliography

Harvard University. (2015). Economists In Cambridge. http://www.fas.harvard.edu/~histecon/visualizing/graphing/economists.html (accessed 4th August 2015).

Hudson, P. ed (2001). *Living Economic and Social History*. Glasgow: Economic History Society.

Jaffe, A., Trajtenberg, M. and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patient citations. *The Quarterly Journal of Economics*, **108**(3): 577-98.

Kuhn, T. ([1962] 1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Laudel, G. (2002). What do we measure by co-authosrhips? *Research Evaluation*, **11**(1): 3-15.

Leydesdorff, L. and Amsterdamska, O. (1990). Dimensions of citation analysis. *Science, Technology, and Human Values*, **15**(3): 305-35.

Lloyd, C. (1995). Economic History and Policy: Historiography of Australian Traditions. *Australian Journal of Politics and History*, **41**(3): 61-79.

Lyons, J., Cain, L. and Williamson, S. ed (2008). *Reflections on the cliometrics revolution: Conversations with economic historians*. New York: Routledge.

Mulkay, M., Gilbert, G. and Woolgar, S. (1975). Problem Areas and Research Networks in Science. *Sociology*, **9**(1): 187-203.

Newman, M. (2010). *Networks: An introduction*. Oxford: Oxford University Press.

Ponds, R., van Oort, F. and Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science*, **86**(3): 423-43.

Shakeosphere. (2015). Shakeosphere: Mapping early modern social networks. http://shakeosphere.lib.uiowa.edu/ (accessed 4th August 2015).

Siler, K. (2013). Citation choice and innovation in science studies. *Scientometrics*, **95**(1): 385-415.

Six Degrees of Francis Bacon. (2015). Six Degrees of Francis Bacon: Reassembling the early modern social network. http://sixdegreesoffrancisbacon.com/ (accessed 4th August 2015).

Wang, C., Rodan, S., Fruin, M. and Xu, X. (2014). Knowledge networks, collaboration networks and exploratory innovation. *Academy of Management Journal*, **57**(2): 484-514.

Whitley, R. (1984). *The Intellectual and Social Organisation of the Sciences*. Oxford: Clarendon.

A Management of Personal Name with Alternate Name and its Searching for Japanese Historical Study

Taizo Yamada
t_yamada@hi.u-tokyo.ac.jp
The University of Tokyo, Japan

Satoshi Inoue
inoue@hi.u-tokyo.ac.jp
The University of Tokyo, Japan

1. Introduction

The basis of historical study or historical understanding consists of collecting historical materials (mainly literature materials such as old documents, old diaries, …), precise reading of the materials, and source criticism. In order to perform them, an identification of a personal name is one of important methods or works, and the researchers of history cannot avoid it. The personal name identification is not simple issue, because there is a diversity in the name representations in the materials. The main representing patterns of the diversity are the follows:

a) Written real or original name
b) Written first name only
c) Written nickname, epithet, or alias
d) Written omitted name
e) Written role name
f) Written using different characters
g) Described Kao (which is a stylized signature or a mark.)

Examples of the representations of "伊集院忠棟 (Ijuin Tadamune)", who is a senior statesman of "島津家 (Shimazu family)" and the relatives of the family and is in the 15th century in Japan, there are "伊集院" (which is his family name), "忠棟" (which is his first name), "幸侃" (which is a nick name, is often appeared in the old documents), "伊右衛門大夫" (which is a nick name), "伊右", "右衛門", "伊大夫" (which are his nick names called by familiar persons), "忠金" (which is his original name). Here is a difficult problem that the various represented names should be identified. In the above example, if you understand the various alternate names can be identified with "伊集院忠棟", the problem is not hard. However, in practice, there are no persons (including researchers of the history) who know and understand all historical persons. For the solution of the problem, we consider that the results of personal name identifications which can be performed by researchers of history should be managed. Furthermore, in the search against historical materials, if the results can be available, the performance of the search can be surely improved compared to simply full-text search.

In the paper we introduce a management method of personal names and the alternate names of the persons and a search method using the managed names. In (Ho, 2015) and (Bol et al., 2015), personal names can be extracted and tagged automated against target documents based on China Biographical Database (CBCB; http://isites.harvard.edu/icb/icb.do?keyword=k16229) as a biographical dictionary. Unfortunately, there are no exhaustive the encyclopedias or dictionaries for the names of Japanese historical persons. Moreover, methods introduced in (Ho, 2015) and (Bol et al., 2015) can be performed better if you can treat a document which is a secondary source like a "地方志 (difangzhi)" in which almost personal names indicates its real name. Most of documents which we treated in the work is primary source and hardly have real names of the person.

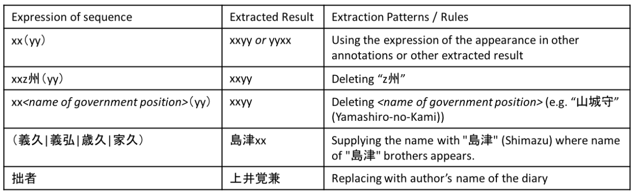
2. Extraction of personal names and alternate names

At first, in order to collecting personal names, we used "上井覚兼日記 (Uwaikakken nikki)" which is a diary of Japanese medieval period (from 1574 until 1586) written by "上井覚兼 (Uwai kakken)" who is a senior statesman of "島津家" of Japan. For the historical study in Kyushu (which is a local area of Japan) or "島津家" in medieval period, the diary is one of important historical materials and Japanese national treasure. The text of the diary has been stored in "The Full-text Database of the Old Japanese Diaries" which has been published by Historiographical Institute, The University of Tokyo. In the text the number of characters is about 1.4 million (for 1777 days; note that there are days which he was not written in the diary). The format of the text is very simply, because the text is just plain text and does not have tags such as XML, TEI. The sample is as follows:

…一、此朝、入来院（重豊）殿太刀を、東郷（重尚）殿次に拳可有之由被申候、御老中よりハ、東郷·祁答院· 入来之事ハ同家にて候間、東郷之次に者根占殿（禰寝重長）太刀を可被召成之由、…

This is a part of the text in the diary of “天正2年8月 1 日” (which indicates A.D. 1574-08-17)”. There are three persons (“入来院 (Irikiin)”, “東郷(Togo)”, “根占 (Nejime)”) in the part. If an alternate name against written name could be solved, the alternate name was added using parenthesis.

The representation can be understood by human (who can read and understand the sentence), but machine can not be solved if the machine doesn’t know or understand the pattern. Due to machine usable, we extracted a written name, an alternate name of the name and the date and we managed the result as a set. As shown in the example, a real name or well-known name is hardly written in the Japanese historical materials, and alternate name added by researcher what indicates real name in most cases. The added name can be used for personal name identification, because this is controlled by the researchers who added the alternate names, and the notation is consistent if the same person. We performed the identification method and could obtain a name pair of 520 sets. In the process of the method, a method of a personal name extraction was needed. We could extract personal names using Machine Leaning method (which consists of an appearance patterns of the names and SVM (Support Vector Machine)). Figure 1 shows examples of the appearance pattern, which an expression of a sequence, an extraction pattern and an extracted result. We used SVM to judge whether the extracted results indicate personal name or not. The SVM results were fed back to the appearance patterns, and we performed the extraction based on appearance patterns and judgment with SVN again. The feedback was preformed several time in the work.



Expression of sequence	Extracted Result	Extraction Patterns / Rules
xx（yy）	xxyy *or* yyxx	Using the expression of the appearance in other annotations or other extracted result
xxz州（yy）	xxyy	Deleting “z州”
xx<*name of government position*>（yy）	xxyy	Deleting <*name of government position*> (e.g. “山城守” (Yamashiro-no-Kami))
（義久|義弘|歳久|家久）	島津xx	Supplying the name with "島津" (Shimazu) where name of "島津" brothers appears.
拙者	上井覚兼	Replacing with author’s name of the diary

Figure 1 Name extraction patterns

We prototyped text search system which supports alternate names using the above constructed personal pairs rather than simple text matching. In the search, for example you queries “忠棟”, then you can obtain the results including “忠棟” as a string, “伊集院忠棟” which is controlled name (well-known name or real name), and alternate names such as “幸侃”, “伊右衛門大夫”, “伊右”, “右衛門”, “伊大夫”, “忠金” (which are mentioned above).

3. Conclusion

Personal name extraction method which we introduced above is useful only target document is “上井覚兼日記 (Uwaikakken nikki)”. In order to extract more generalization, preparing a pattern suitable to each historical material is necessary. We constructed a database which can be stored the personal name pairs. Currently we also have been collecting personal names from other texts and databases, and storing it in the database. The data in the database indicates the results as an identification of personal name. The data can assist to identify personal names in a material which reading comprehension has been not yet done. We expect that the method can be useful to progress of the study of Japanese history.

Bibliography

Ho, H. I. (2015). MARKUS – A Fundamental Semi-automatic Markup Platform for Classical Chinese. *Proceedings of the 2015 International Conference on Digital Humanities.*

Bol, P., Liu, Ch.-L., and Wang, H. (2015). Mining and Discovering Biographical Information in Difangzhi with a Language-Model-based Approach. *Proceedings of the 2015 International Conference on Digital Humanities.*

Computation-Aided Analysis on Film Credits

Li Yang
yangll@lafayette.edu
Lafayette College, United States of America

Weijia Xu
xwj@tacc.utexas.edu
Texas Advanced Computing Center, University of Texas at Austin, United States of America

Introduction

This project is centrally concerned with using digital computation to analyze historical and comparative data of film credits in order to study the evolution of film production. Our hypothesis is that the elements of film production, as credited by the film itself, have a direct correlation with the historical and cultural background of film production. We want to identify the specific connection with each query, hence contributing to the understanding of the culture of film production in general.

In the introduction to *The Production of Culture and the Cultures of Production*, Paul du Gay points out that culture and economy are mutually constitutive in the present day. Not only are most of popular cultural products manufactured by organized industries, but the production activities themselves also manifest cultural values (Gay, 1997). Our project is theoretically oriented toward the latter, the cultures of production. Although relatively less studied than the production of culture led by the theories of the Culture Industry (Horkheimer & Adorno, 1997), significant scholarship has been conducted to discern the cultural patterns of the film and television productions, such as John Thornton Caldwell's book, *Production Culture: Industrial Reflexivity and Critical Practice in Film and Television* (Caldwell, 2008). The primary data Caldwell and other researchers used are trade publications, interviews, and field observations. Sharing the same theoretical inclination of Caldwell, our research breaks the methodological new ground by analyzing the most conspicuous records of film production—the credits shown on film itself.

Influenced by fiction publication and theater programs in earlier years, film crediting convention has evolved throughout the history of cinema. Today, it is typical for an American film to open with a title sequence and end with a longer rolling credit crawl. The end credits for big production blockbusters are relatively lengthy, featuring both above-the-line categories such as director, producer, and cast, and below-the line categories such as production assistants, camera operators, electricians, special effects specialists, sound editors, etc. As records of film production, film credits contain a plethora of vital information, disclosing not only involved labor (personnel), and corporations (financiers, production companies, and distribution companies), but also procedures (e.g. special effects units and location shootings), and technologies (e.g. Panavision cameras and Chapman camera cranes).

The scholarly attention to film credits has steadily increased in recent years. Many focus on the credits' aesthetic and philosophical values (de Mourgues, 1993; Moinereau, 2009; Tylski, 2009). Some use film credits as an evidence to study the Hollywood star system and labor hierarchy (Clark, 1995; Chisholm, 2000; Carman, 2008). Will Straw's article "Letter of Introduction: Film Credits and Cityscapes" represents a significant contribution by linking the design of the title credit sequence to the socio-historical condition of film viewing – the development of city life after World War II (Straw, 2010). Two recent dissertations also offer interesting insights (Allison, 2001; Crawford, 2013). Allison's research is statistical in nature but only confined to opening title sequences. Crawford's dissertation presents multiple perspectives (aesthetic, legal, and industrial), convincingly justifying the importance of film credits.

This project focuses on actually accessing and analyzing film credits to generate insights into the evolution of film production. More specifically, we propose to employ large-scale computation to perform comparative analysis on film credits. Digital computation is necessary to carry the study of film credits to another level since physically examining the credits even for one film is a daunting task. The final crawl of *Dark Knight* (2008), for example, runs for 7.5 minutes and contains hundreds of names and entities. The listed order and naming of some categories also differ from film to film. Besides increasing efficiency by substituting for human labor, digital computation also invites a large quantity of research subjects. These methods will potentially discover patterns across an extended period of time as well as across national and genre boundaries. We are confident that the potential of film credits as written records will be fully tapped with the aid of digital computation, and our research project will represent a major contribution to film studies.

Our project is mapped into two stages. For the first stage, we will pull available digital records of American films on the Internet to perform some statistical analysis. The second stage will see the expansion of selected works from American films to Chinese films. The Chinese film industry is chosen because of its dramatic development trajectory in the last 30 years. The industry was resuscitated in the 1980s from its near death in the Cultural Revolution (1966–1976) (zero output for three years), and only found itself slipping into financial crises throughout the 1990s in its attempt to balance political and commercial imperatives. Just when the industry was on the brink of collapse at the dawn of the new millennium, it picked itself up with full-scale capitalization while still maintaining substantial degree of governmental control. Charting an extraordinary growth curve to the present day, the Chinese film market has become the second largest in the world since 2012, and is on track to overtake the United States in 2017 or 2018 if the current growing pattern sustains. Tracing the information listed as credits over time and comparing to that of Hollywood will demonstrate the trajectory of both local changes and external influences in the development of Chinese commercial cinema. Because the digital records of Chinese film credits are not as readily available as their American counterparts, we will employ video-capture as well as optical character recognition technology to establish source data sets.

Preliminary Findings

In this conference paper we will present some of our preliminary findings from the first stage. We have developed a software tool to aggregate and analyze records from International Movie Database (IMDB) for analysis. The application is implemented in Java and can parse IMDB records, extract named entities, conduct statistical and association analysis. One of our queries is the number of people listed under major credit categories as documented by the IMDB. We selected 20,690 titles since 1900 with at least one director and at least 10 cast members (actors and actresses). The average number of crew for each film over years is shown in Table-1. For 1900 to 2010, the film statistics are aggregated every 10 years. For example, "1900 to 1910" refers to films released between Jan. 01, 1900 and Jan. 01, 1910. Figure 1 shows the growth of cinematographer, composer, costume-designer, director, editor, miscellaneous workers, producer, production-designer, and writer categories. And Figure 2 shows growth in the number of actors and actresses.

	ACTOR	ACTRESS	CINEMATOGRAPHER	COMPOSER	COSTUMER	DIRECTOR	EDITOR	MISC	PRODUCER	PRODUCTION	WRITER	Movie count
1900 to 1910	9.57	6.57	0.57	0.29	0.00	2.14	0.00	0.14	1.14	0.14	2.00	7
1910 to 1920	9.67	4.96	0.94	0.08	0.05	2.12	0.08	0.24	0.76	0.04	1.73	347
1920 to 1930	10.56	4.87	1.29	0.40	0.15	2.14	0.37	0.68	0.83	0.07	2.61	492
1930 to 1940	19.40	6.23	1.31	0.91	0.26	2.16	0.81	0.96	1.01	0.28	3.16	971
1940 to 1950	21.89	7.64	1.26	1.01	0.33	2.24	1.06	1.46	1.14	0.43	3.20	613
1950 to 1960	19.52	8.72	1.19	0.99	0.34	2.41	0.95	1.60	1.26	0.48	3.18	863
1960 to 1970	17.37	8.52	1.15	0.91	0.35	2.60	0.96	1.31	1.27	0.43	3.00	1045
1970 to 1980	16.59	7.90	1.0	0.84	0.29	2.50	0.82	1.97	1.49	0.30	2.60	1301
1980 to 1990	18.21	9.51	0.91	0.84	0.34	2.71	0.86	3.69	1.98	0.38	2.75	1559
1990 to 2000	18.47	11.06	0.95	0.83	0.38	2.96	1.04	5.86	2.90	0.34	2.89	2127
2000 to 2010	15.89	8.84	1.18	0.90	0.30	2.60	1.35	5.54	3.68	0.32	2.37	5458
2010 to 2015	15.47	8.26	1.29	0.85	0.25	2.55	1.36	4.35	4.50	0.31	2.37	5655
after 2015	18.04	9.50	1.17	0.65	0.30	3.04	1.11	5.37	5.57	0.37	2.47	252

Table 1 Average number of crew per film from 1900 to 2015

From Figure 1, we can see the dramatic increase in the number of credited producers and miscellaneous workers (below-the-line film crew) over time. After years of steady increase, the number of producers took off around 1975, the same year that saw the release of modern blockbuster prototype *Jaws*. Our findings put a direct link between the spreading of the blockbuster business model and the number of producers on board. The miscellaneous worker category displays an even steeper rise. Interestingly, the time of dramatic increase coincides with that of producers – around 1975, but its down size in the early 1990s has no parallels. This graph pinpoints key historical moments for further investigation in the area of union activities, film labor division and representation. Finally from Figure 2 we can see that the gap between male and female cast has remained steady over a century. The biggest gap, however, does not appear in recent years when superhero films reign at the box office – probably thanks to an increased awareness of gender equality – but in the period of 1925 to 1945. Our findings invite further scrutiny into the pre-Paramount Decree (1948) years to study gender dynamics in Hollywood. From this perspective, the conventionally-defined golden age of Hollywood studio era (1927–1960) (Bordwell et al., 1985) may not be as seamlessly continuous as once conceived.

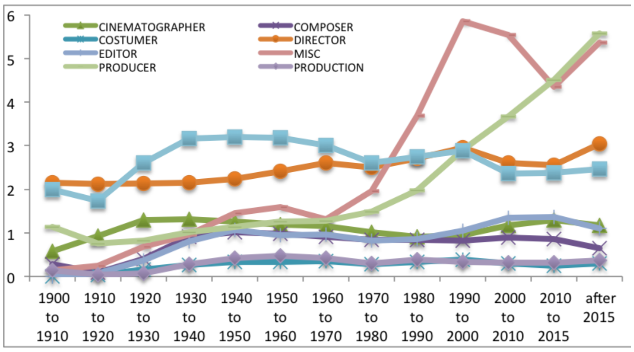


Figure 1 Average number of crews excluding actors and actress per film

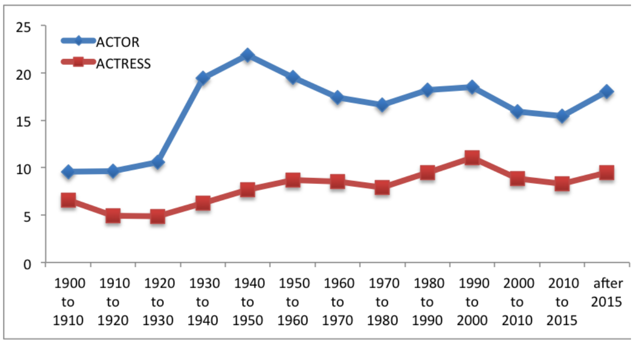


Figure 2 Average number of actors and actress per film

Summary

In summary, these preliminary findings illustrate the mechanism of using computation to analyze the credits of a large number of films across historical periods. The inquiries described above all fall into the category of the "personnel." There are three other categories for future considerations: companies, procedures, and technologies. We believe that the computation-aided study of film credits will detect patterns or gaps, which will serve as important clues and evidences to analyze the cultural dynamics of film production.

Bibliography

Allison, D. (2001). *Promises in the Dark: Opening Title Sequences in American Feature Films of the Sound Period*. Dissertation. ProQuest Dissertations Publishing.

Bordwell, D., Staiger, J. and Thompson, K. (1985). *The Classical Hollywood Cinema: Film Style and Mode of Production to 1960*. New York, NY, USA: Columbia University Press.

Caldwell, J. T. (2008). *Production Culture: Industrial Reflexivity and Critical Practice in Film and Television*. Durham, NC: Duke University Press.

Carman, E. S. (2008). Independent Stardom: Female Film Stars and the Studio System in the 1930s. *Women's Studies: An Inter-disciplinary Journal*, **37**(6): 583–615.

Chisholm, A. (2000). Missing Persons and Bodies of Evidence. *Camera Obscura*, **15**(1): 122–61.

Clark, D. (1995). *Negotiating Hollywood: The Cultural Politics of Actors' Labor*. Minneapolis, MN: University of Minnesota Press.

Crawford, J. A. M. (2013). *Film Credit*. Dissertataion. ProQuest Dissertations Publishing.

de Mourgues, N. (1993). *Le Générique de Film*. Paris: Méridiens Klincksieck.

Gay, P.d. (1997). *Production of Culture/Cultures of Production*. London: Sage.

Horkheimer, M. and Adorno, T. W. (1997). *Dialectic of Enlightenment*. Translated by J. Cumming. New York: Continuum.

Moinereau, L. (2009). *Le Générique de Film: de la Lettre à la Figure*. Rennes: Presses Universitaires de Rennes.

Straw, W. (2010). Letters of Introduction: Film Credits and Cityscapes. *Design and Culture: The Journal of the Design Studies Forum*, **2**(2): 155–65.

Tylski, A. (2009). *Le Générique au Cinéma: Histoire et Fonctions d'un Fragment Hybride*. Lille: Presses Universitaires du Mirial.

From Language Revolution to Revolutionary Language

Ching-Syang Jack Yue
csyue@nccu.edu.tw
National Chengchi University, Taiwan, Republic of China

Li-Hsing Ho
lillianlhho@gmail.com
National Tsing Hua University, Taiwan, Republic of China

Wen-Huei Cheng
wenhuei_cheng@yahoo.com.tw
National Chengchi University, Taiwan, Republic of China

In his insightful book *The Language of the Third Reich*, Jewish linguist Victor Klemperer pointed out, "One tends to understand Schiller's distich on a 'cultivated language which writes and thinks for you' in purely aesthetic and, as it were, harmless terms [...] But language does not simply write and think for me, it also increasingly dictates my feelings and governs my entire spiritual being the more unquestioningly and unconsciously I abandon myself to it." Thus he observed and detailed the language change during and after the rise of Hitler and Nazism, kept an invaluable record for study of the impact of language on the mind.

Klemperer was by no means the only scholar who was alert to the influence of language. In Chinese context, when May Fourth intellectuals followed the flow of turning the classical Chinese into something more colloquial in late Qing period and advocated a new literature/cultural movement, which was at its heart a language revolution, the idea under the action was that one cannot separate the language habits from the operation of mind. They started the vernacular movement in magazines like *the New Youth*, and hoped by modernizing language, they would empower the mind of the nation, modernize the culture, and eventually really modernize China, a task that 1911 revolution apparently did not fully succeed. The language revolution in early Republic was closely studied by many scholars. Our team also contributed by bringing in digital humanity methods, finding effective ways to tell the classical Chinese from modern Chinese by computer. It was when we analyzed the data of *The New Youth Magazine* that we noticed something rarely discussed by modern scholars: the language changed significantly in Volume 8 of the magazine, published on the eve of the rise of Chinese Communist Party.

According to our study, by the end of Volume 7 (published in 1920), modern vernacular Chinese had almost completely replaced classical Chinese as the main written language. However, the language of Volume 8 seems to be a new breed, deeply influenced by the translation of Soviet language and full of political jargons. It seems that right after success of language revolution, *The New Youth Magazine* immediately change the language again and promote a new kind of "revolutionary language" to advocate the Communism before the Chinese Communist Party was founded. Was there someone like Klemperer to observe the language change before the political turmoil? It called for further study. What we are doing here is to analyze the language turn in the late period of *The New Youth Magazine*. For comparison, we draw in two other materials. First, the editorials of *United Daily News* from 1951 to 1960, inherited the May Forth legacy and published in Taiwan before Taiwanese Modernist movement in 60s and Indigenized movement in 80s changing the written language violently. Second, the essays from the Chinese Communist Party's *People's Daily* from 1971 to 1989, published in mainland China before the Chinese economic reform.

There are two types of data: one is structured data (with a high degree of organization) and the other is unstructured data. Most of the textual data are unstructured and quantifying them often requires certain knowledge about the application domain. We need to create a relational structure for the textual data before plugging classification methods. In particular, we use the notion of Exploratory Data Analysis (or EDA, proposed by famous statistician J.W. Tukey in 1977) to evaluate potential variables which can differentiate the language styles of Volume 7 and Volumes 8~11 in *New Youth Magazine*. According to our previous study of writing style for Volumes 1~7, we found that the number of words, the number of different words (or vocabularies) and their distribution, in addition to function words in classical and modern Chinese, can be used to distinguish Chinese writing styles. We also include species diversity indices, such as Simpson index and entropy (or Shannon index).

Since there are more than 30 variables, we also consider data reduction methods, such as principle component analysis, to include fewer variables. Then, we apply classification methods (logistic regression) to judge the style of an article is close to Volume 7 or Volumes 8~11. Also, to avoid over parameterization (i.e., using too many unnecessary variables), we use cross validation to check whether the model is stable. The regression model is first built based on training data and then applied to the testing data. The fitting accuracies of training data and testing data are recorded separately, and these two numbers of accuracy should be close if the model is stable. For every simulation run, we randomly separate the training data and testing data into proportions of 90% and 10%, respectively. Based on 100 simulation runs, the regression model is fairly stable since it has very similar fitting accuracy (and small standard errors) for training and testing data.

Next step, we apply the constructed regression model to the articles from two newspapers, *United Daily News* and *People's Daily*. Both newspapers have about 500 articles and Table 1 shows their classification results. The *United Daily News* is published in Taiwan and the Soviet Union should have little influence on this newspapers. As expected, only 10% articles are classified to the group of "Revolutionary Language." On the other hand, more than 50% of articles from *People's Daily* are classified to "Revolutionary Language." This matches to our anticipation, since the articles from *People's Daily* are published in 1971~1989, in which period China had close link with the Soviet Union.

No. of Article	**Classified to "Revolutionary Language"**	**No. of Article**	**Proportion**
United Daily News	550	57	10.36%
People's Daily	534	308	58.68%

Table 1. Classifications of *United Daily News* and *People's Daily*

Similar to the cross-validation for the articles from the *New Youth magazine*, we can fit the regression model to the articles from two newspapers year by year. Figure 1 shows the yearly classification results. It seems that the results of *United Daily News* are fairly stable and the yearly results lie around the average (red dotted line). The fitting results of *People's Daily* somehow show an inconsistence pattern. The average proportion of "Revolutionary Language" is 76.3% for the articles in 1971~1977 and is 39.2% in 1978~1989 (the overall average is the red dotted line).

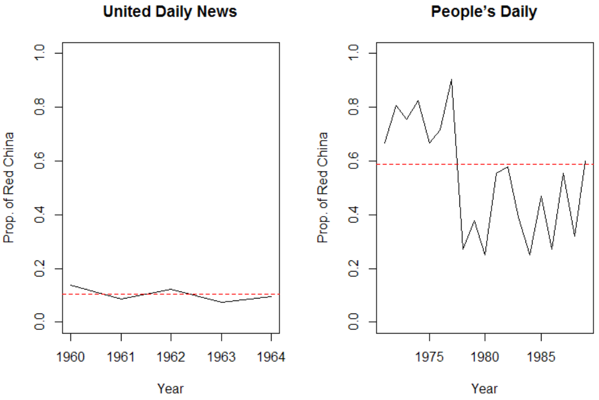


Figure 1. Yearly Classifications of *United Daily News* and *People's Daily*

The yearly classification results of *People's Daily* are interesting and very encouraging. The first stage of China's reform and opening is between 1978~1989, under the formal leader Deng Xiaoping. China started to open to the outside world gradually since 1978, and it also triggers the economic blooming of China at the end of 20th century. Figure 1 suggests a similar story. The proportion of "Revolutionary Language" (under the influence of Soviet Union) articles is larger before 1978 and has a radical change ever since 1978 until 1989, coincide with the years of reform and opening in modern China.

2015 marks the 100th anniversary of The *New Youth Magazine*. By employing the digital humanity methods and focusing on the language change of the latest few volumes that so far rarely discussed by scholars, we hope to make some new contribution to the study of both The *New Youth Magazine* and the writing style of modern Chinese language. Our study shows that the styles of articles from *United Daily News* are close to that of Volume 7 in *New Youth Magazine*, while those from *People's Daily* have about equal probability for being classified to Volume 7 or to Volumes 8~11. This implies that the Chinese writing styles of Taiwan and mainland China are different and the writing style of mainland China is likely to be influenced by the Soviet language.

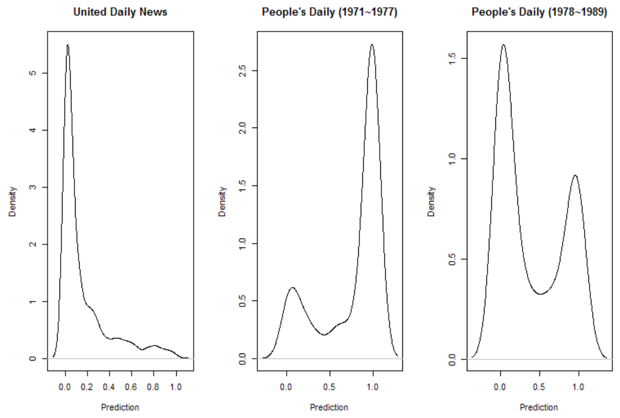


Figure 2. Predicted Results of *United Daily News* and *People's Daily*

Also, the change of "Revolutionary Language" in 1978 seems to indicate a new possible writing style. Judging from the proportion of "Revolutionary Language," those of 1978~1989 lie between "Revolutionary Language" and *United Daily News,* and it seems that the writing style of China turned into a new direction in 1978~1989 (namely, Reform China). Figure 2 shows the predicted results (after kernel smoothing) of all articles, with a value near 1 indicating style closer to "Revolutionary Language" and vice versa. If we treat *United Daily News* and 1971~1977 *People's Daily* as two extremes, then 1978~1989 *People's Daily* is somewhat in between (or a mixture of) these two extremes. It would be interesting to explore the writing style in China since 1978. For example, we can further compare the writing styles of China according to the following three periods, 1978~1989 (first stage), 1989~2002 (economic blooming), and 2002~today (modern China).

Bibliography

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition, Springer Series in Statistics.

Ho, L., Yue, C. J., and Cheng, W. (2014). From Classical Chinese to Modern Chinese: A Study of Function Words from New Youth Magazine. *Journal of the History of Ideas in East Asia*, 7: 427-54.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Posters

Content Based Social Network Analysis of Reşat Nuri Güntekin's Letters

Sumeyye Akca
sumeyyesakca@gmail.com
Hacettepe University, Turkey

Muge Akbulut
mugeakbulut@gmail.com
Yıldırım Beyazıt University, Turkey

Reşat Nuri Güntekin is one of the most important authors in Turkish literature, who gave a direction to the Turkish literature with his many novels and essays. He also conducted significant bureaucratic tasks in the early years of the Turkish Republic. He left substantive materials which have historical and literary value, considering his lifetime. The letters he wrote to his wife give us important information about this period. In addition, these letters are also of capital significance in terms of their evincing the historical events and social life of early period of Turkish Republic through the eyes of an author and statesman. The letters also constitute alternative historical as well as official information resources.

The letters began to be written to his wife, Hadiye in 1927. Although 62 letters were written with their dates affixed, there are 22 letters without dates. The letters were written in Ottoman Turkish. They were transcribed into the Latin characters and published. In order to discover the relationships between events and people in the text, we visualize the letters by using social network analysis and content analysis methods.

Social network analysis method is used to visualize the connections between communities whose existing connections are not easy to be perceived and modelled, through describing the network structures. Social network analyses are important in terms of visualizing the complexity in a simpler way. The fact that the letters Reşat Nuri Güntekin sent to his wife were from different cities and about important people of the period makes them suitable for a social network analysis. With the social networks obtained from the archive of letters, the closeness between the historical figures and the indirect effects to each other can be detected and the historical inferences can be made by examining these connections.

Content analysis method is the process of summarizing the content of the written information and the messages contained in them. This method has been used frequently in social sciences and typically preferred for the analysis of written texts such as books, letters, newspapers etc.

Primarily the descriptive statistics about the content of the letters are presented and then the findings from the content analysis on the messages are revealed. In this paper, the big picture and the details are explained through social network analysis method and the connections and relationships which are indirect, are made apparent.

Relationship between people and the citations etc. are mostly suggested clearly in social network analysis studies. However the relationship between the people in the mentioned letters is not clear. Therefore it is determined in the study based on the proximity and frequency of the names in the text. Proximity-based method is based on the names similar to each other in the letters. Within the scope of the study, CiteSpace software was used for visualization.

Engaging Students in Digital Literary Analysis: GALGO (Golden Age Literature Glossary Online), a Social Semiotic Platform

Nuria Alonso Garcia
nalonsog@providence.edu
Providence College, United States of America

Alison Caplan
acaplan@providence.edu
Providence College, United States of America

This poster showcases a digital teaching application that approaches the study of language and literary texts from a social semiotic perspective and represents an innovative pedagogical model for world language and literature classes. The Golden Age Literature Glossary Online, known by the acronym *GALGO*, consists of an online glossary of select keywords, from canonical texts of Golden Age Spanish literature, whose multiple connotations illuminate important linguistic and social concepts of the 16th and 17th centuries. *GALGO* incorporates British cultural historian Raymond Williams' methodology in his *Keywords: A Vocabulary of Culture and Society*: namely, identifying problem-laden words or "keywords," charting their distinct usages across texts, and reflecting critically on clusters of associated words. Implicit in William's keyword analysis is a social semiotic theory of language that takes as its starting point the observation that meanings "are created by the social system and are exchanged by the members in the form of text" (Halliday, 1978, p.141) and find in literary texts their fullest creative expression (Lotman, 1990). Contextualizing poetic language in its particular space-time, therefore, reveals the linguistic codes derived from the culture in which the work has been produced.

GALGO seeks to instantiate the theoretical construction that embeds the semantic configurations of a literary text simultaneously in the cultural environment, the linguistic system, and the social system. The computer is an ideal semiotic machine to expose these multiple interwoven strands of meaning at once. Applying the structure of M.A.K. Halliday's theoretical model in *Language as social semiotic*, *GALGO*'s interpretive apparatus allows the user to identify the *field of discourse* or social function of each instance of a specific keyword and determine its distributional profile[1](DP) with respect to underlying semantic relationships. The platform is capable of presenting a list of clusters, word groupings based on semiotic affinity, along with their fuller contexts in the works. Simultaneously, within the cluster analysis, *GALGO* also prompts an interpretation of the *tenor of discourse*, highlighting sociological variables connected to class status, gender role and racial category that refine the text's meaning from an interpersonal perspective.

We envision the architecture of the platform to not only be suitable for our corpus of 16th and 17th century Spanish works, but rather potentially abstracted in such a way that any set of words could be glossed from any collection of works. *GALGO* offers students the opportunity to explore a wide variety of views and re-combinations of words and definitions in order to engage more effectively with primary sources that are written in a language, discipline, or time period that is "foreign" to an undergraduate. In this, its third prototype, *GALGO* suggests untapped potential for computer-assisted textual analysis and would now benefit from additional critical feedback offered by the larger digital humanities community.

Participants will have the chance to engage with the platform during the poster session and share their feedback and experiences with the authors:

- How user-friendly is GALGO' s interface?
- How successful is GALGO in performing its intended tasks (cultural semiotic analysis, sociolinguistic research, student engagement in reading)?
- How can GALGO be improved?
- How do others engage students with new media methods?
- How suitable is GALGO for texts from other disciplines?

Bibliography

Halliday, M. A. K. (1978). *Language as social semiotic. The social interpretation of language and meaning*. Baltimore: University Park Press.

Lotman, Y. M. (1990). *Universe of the mind. A semiotic theory of culture*. London: I.B. Taurus & Co.

Rubenstein H. and Goodenough J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**(10): 627-33.

Schütze, H. (1992). Dimensions of meaning. *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pp. 787-96.

Notes

[1] According to the Distributional Hypothesis, the distributional profile (DP) of a word is determined by the strength of association of the word with co-occurring words in the text (Rubenstein & Goodenough, 1965, Schütze, 1992)

The DEFC-App: A Web-based Archaeological Data Management System for 'Digitizing Early Farming Cultures'

Peter Andorfer
Peter.Andorfer@oeaw.ac.at
Austrian Centre for Digital Humanities, Austrian Academy of Sciences

Edeltraud Aspöck
Edeltraud.Aspoeck@oeaw.ac.at
Institute for Oriental and European Archaeology, Austrian Academy of Sciences

Matej Ďurčo
Matej.Durco@oeaw.ac.at
Austrian Centre for Digital Humanities, Austrian Academy of Sciences

Anja Masur
Anja.Masur@oeaw.ac.at
Institute for Oriental and European Archaeology, Austrian Academy of Sciences

Ksenia Zaytseva
Ksenia.Zaytseva@oeaw.ac.at
Austrian Centre for Digital Humanities, Austrian Academy of Sciences

Starting Point

The project objective of **D**igitizing **E**arly **F**arming **C**ultures (DEFC) is the standardization and integration of research data from sites and finds from the Neolithic and Copper Age (7000–3000 BC) located in Greece and Western Anatolia. These datasets are based on digital and analog resources of research projects of the research group Anatolian Aegean Prehistoric Phenomena (AAPP) at the Institute for Oriental and European Archaeology (OREA) of the Austrian Academy of Sciences.

Greece and Western Anatolia are two neighbouring and archaeologically closely related regions. They are, however, usually studied in isolation from each other and have therefore developed different terminologies and chronologies. Direct results of this de facto separation are not only huge amounts of fragmented research data but also several different models and standards for ordering and describing more or less the same kind of data. To pose and answer archaeological research questions concerning the whole territory, the information must be harmonized.

The aim of the DEFC project is now to harmonize the existing data, to digitize analog resources and make metadata available to facilitate access and reuse this data. To achieve those goals an archaeological data management system is needed.

Data model and Application

The particular requirements to the data model are to reflect the high granularity of the archaeological data structure which correlates on different levels to the excavation process workflow, geographical location, chronological periodization and at the same time to keep the complex relationships between the data objects. After an evaluation of already existing solutions for managing (archaeological) data (e.g. Microsoft Access, Arches Project) it turned out that those were not comprehensive enough for modeling and capturing the very heterogeneous datasets the DEFC project is confronted with. Therefore the development of a more customizable application to collect, standardize, analyze and visualize archaeological data was necessary.

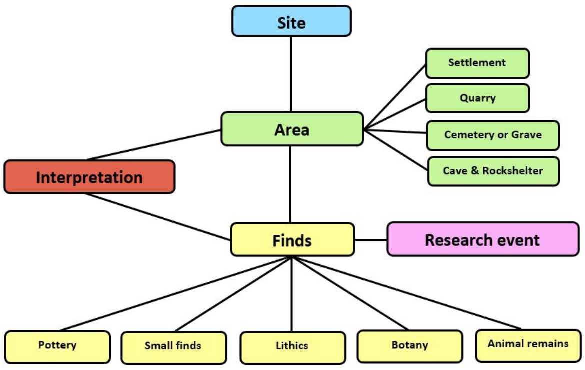
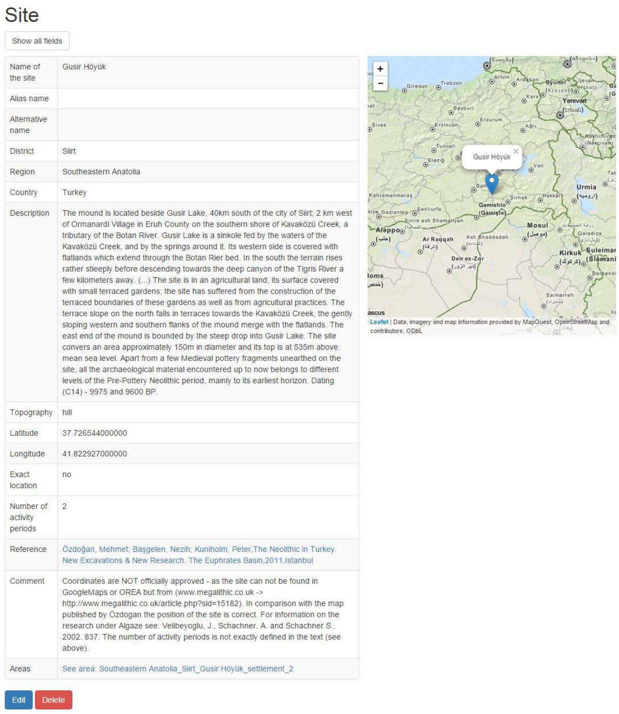


Figure 1. Simplified data model

To meet the needs of researchers a clear conceptual data model based on archaeological objects relationships has been defined with the following main model classes:

- Site (location where research took place/observations were made)
- Research Event (project and type of archaeological research that was carried out)
- Area (particular part of the site, defined by its geolocation, period, as well as its type)
- Finds (artefacts, animal and plant remains found)
- Interpretation (archaeologist's interpretation of areas/finds etc.)

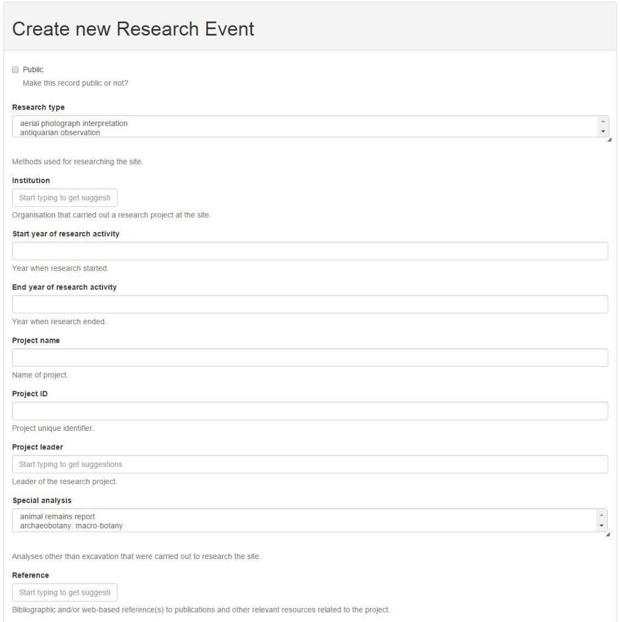
Each of those classes is defined through several properties, most of them linked to a carefully curated set of controlled vocabulary.The DEFC-App is based on the Python web framework Django. As one of the application's design principles is to keep things as simple as possible, the application tries to leverage Django´s built-in generic functionality as far as possible. The application's web interface is based on Bootstrap. Client-side scripting, which is needed for a better user guidance and enabling a more responsive data querying and presentation, is implemented with JavaScript, jQuery, Tablesorter and Leaflet.



Site

Name of the site	Gusir Höyük
Alias name	
Alternative name	
District	Siirt
Region	Southeastern Anatolia
Country	Turkey
Description	The mound is located beside Gusir Lake, 40km south of the city of Siirt; 2 km west of Ormanardi Village in Eruh County on the southern shore of Kavaközü Creek, a tributary of the Botan River. Gusir Lake is a sinkole fed by the waters of the Kavaközü Creek, and by the springs around it. Its western side is covered with flatlands which extend through the Botan Rier bed. In the south the terrain rises rather stieeply before descending towards the deep canyon of the Tigris River a few kilometers away. (...) The site is in an agricultural land, its surface covered with small terraced gardens; the site has suffered from the construction of the terraced boundaries of these gardens as well as from agricultural practices. The terrace slope on the north falls in terraces towards the Kavaközü Creek, the gently sloping western and southern flanks of the mound merge with the flatlands. The east end of the mound is bounded by the steep drop into Gusir Lake. The site convers an area approximately 150m in diameter and its top is at 535m above mean sea level. Apart from a few Medieval pottery fragments unearthed on the site, all the archaeological material encountered up to now belongs to different levels of the Pre-Pottery Neolithic period, mainly to its earliest horizon. Dating (C14) - 9975 and 9600 BP.
Topography	hill
Latitude	37.726544000000
Longitude	41.822927000000
Exact location	no
Number of activity periods	2
Reference	Özdoğan, Mehmet; Başgelen, Nezih; Kuniholm, Peter,The Neolithic in Turkey. New Excavations & New Research. The Euphrates Basin,2011,Istanbul
Comment	Coordinates are NOT officially approved - as the site can not be found in GoogleMaps or OREA but from (www.megalithic.co.uk -> http://www.megalithic.co.uk/article.php?sid=15182). In comparison with the map published by Özdogan the position of the site is correct. For information on the research under Algaze see: Velibeyoglu, J., Schachner, A. and Schachner S., 2002. 837. The number of activity periods is not exactly defined in the text (see above).
Areas	See area: Southeastern Anatolia_Siirt_Gusir Höyük_settlement_2

Figure 2. Site details page



Create new Research Event

□ Public
Make this record public or not?

Research type
aerial photograph interpretation
antiquarian observation
Methods used for researching the site.

Institution
Start typing to get suggesti
Organisation that carried out a research project at the site.

Start year of research activity
Year when research started.

End year of research activity
Year when research ended.

Project name
Name of project.

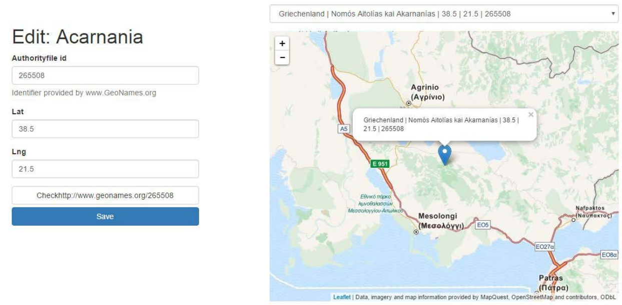
Project ID
Project unique identifier.

Project leader
Start typing to get suggestions
Leader of the research project.

Special analysis
animal remains report
archaeobotany: macro-botany
Analyses other than excavation that were carried out to research the site.

Reference
Start typing to get suggesti
Bibliographic and/or web-based reference(s) to publications and other relevant resources related to the project.

Figure 3. Create new Research Event page



Edit: Acarnania

Authorityfile id

265508

Identifier provided by www.GeoNames.org

Lat

38.5

Lng

21.5

Checkhttp://www.geonames.org/265508

Save

Griechenland | Nomós Aitolías kai Akarnanías | 38.5 | 21.5 | 265508

+
−
Agrinio
(Αγρίνιο)
Griechenland | Nomós Aitolías kai Akarnanías | 38.5 | 21.5 | 265508
A5
E 951
Εθνικό πάρκο λιμνοθαλασσών Μεσολογγίου-Αιτωλικού
Mesolongi
(Μεσολόγγι)
EO5
Nafpaktos
(Ναύπακτος)
EO27α
EO8α
Patras
(Πάτρα)
Leaflet | Data, imagery and map information provided by MapQuest, OpenStreetMap and contributors, ODbL

Figure 4. Referencing Geonames

Development and Upcoming tasks

The project aims to integrate open access resources by using Web APIs and Linked Data practices. At the time being Geonames referencing is implemented for the archaeological locations and provided via a user interface. Hence the fetched Geonames IDs are stored within the database to later be linked to the Pelagios project.

The bibliographic data formerly stored in proprietary formats MS-Access and AskSam was imported to a Zotero library and linked to the DEFC-Database so that every reference record in DEFC redirects to a Zotero library record, where the entire bibliography can also be explored.

To make the published data available 'open access' for further reuse in research, a REST-API (Django REST framework) was implemented along with the web user interface for querying and exporting data.

The outlook of the project is to turn the data into Linked Open Data and make it available via a SPARQL endpoint. Moreover, the thesaurus consisting of hierarchically structured archaeological data units (respectively the aforementioned controlled vocabulary) has been partially mapped to the CIDOC CRM ontology and will later be mapped to the SKOS schema. This will, overall, enhance the quality of the RDF data in the future.

Conclusion

The development of the DEFC-App and its underlying data model could be understood as a very common use case in the broad field of digital humanities as it involves a tight cooperation between archaeologists, data analysts and developers.

Bibliography

Arches project. [Online] Available from: http://archesproject.org/ [Accessed 4 March 2016].

Christian Bach. *Tablesorter*. [Online] Available from: http://tablesorter.com/docs/ [Accessed 4 March 2016].

DEFC-App. [Online] Available from: http://defc.digital-humanities.at/ [Accessed 4 March 2016].

Django Software Foundation and individual contributors. *Django*. [Online] Available from: https://www.djangoproject.com/ [Accessed 4 March 2016].

Django REST framework. [Online] Available from: http://www.django-rest-framework.org/ [Accessed 4 March 2016].

Geonames. [Online] Available from: http://www.geonames.org/ [Accessed 4 March 2016].

Pelagios: Enable Linked Ancient Geodata In Open Systems. [Online] Available from: http://pelagios-project.blogspot.co.at/p/about-pelagios.html [Accessed 4 March 2016].

Vladimir Agafonkin. *Leaflet*. [Online] Available from: http://leafletjs.com/ [Accessed 4 March 2016].

Zotero. [Online] Available from: https://www.zotero.org/ [Accessed 4 March 2016].

Pundit. Semantic Annotation for Digital Humanities

Giulio Andreini
andreini@netseven.it
Net7 S.r.l., Italy

Francesca Di Donato
didonato@netseven.it
Net7 S.r.l., Italy

Danilo Giacomi
giacomi@netseven.it
Net7 S.r.l., Italy

Enrico Giusti
giusti@netseven.it
Net7 S.r.l., Italy

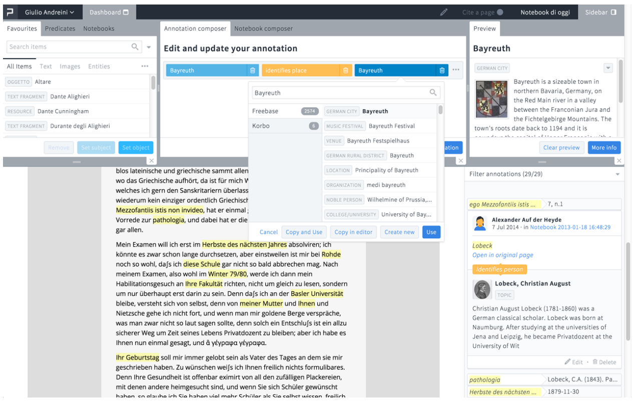
Raffaele Masotti
masotti@netseven.it
Net7 S.r.l., Italy

Students and researchers are used to study on books and printed articles, underlining and taking notes on the text itself in the meantime.

Pundit (http://thepund.it/) allows to perform the same actions on any web page, being it an online magazine, a blog or a digital library. In fact, Pundit (Morbidoni et al., 2015) is an open source suite of applications that allows users to build semantic annotations with different levels of expressivity on web pages, collaborating in the meantime with others. The knowledge base created by annotations can be reused inside the Pundit applications or by external third party projects.

Pundit implements a client-server architecture and is made up of several components which interact with each other, but, if needed, are able to work independently:

- **Pundit Annotator**: a simple and lightweight annotator tool that allows to highlight and comment text in web pages with ease. This tool is intended for general users, students and journalists.
- **Pundit Annotator Pro**: an advanced tool for web annotation that allows to create semantic annotations (built by one or more triples) using text fragments, web pages, Cultural Heritage Objects or Linked Data entities. This tool is intended for scholars and researchers.
- **Pundit Annotations Manager**: the Annotation Manager is a web application that allows users to review and manage their annotations as well as to export them in different formats.
- **Pundit Server**: it is where all annotations are stored in a graph format. The data model of annotations is an extension of the Web Annotation Data Model standard (http://www.w3.org/TR/2014/WD-annotation-model-20141211/), defined by the Web Annotation Working Group of W3C.



Bayreuth is a sizeable town in northern Bavaria, Germany, on the Red Main river in a valley between the Franconian Jura and the Fichtelgebirge Mountains. The town's roots date back to 1194 and it is

blos lateinische und griechische sammt allen
wo das Griechische aufhört, da ist für mich W
welches ich gern den Sanskritariern überlass
wiederum kein einziger ordentlich Griechisch
Mezzofantiis istis non invideo, hat er einmal
Vorrede zur pathologia, und dabei hat er die
gar allen.

Mein Examen will ich erst im Herbste des nächsten Jahres absolviren; ich könnte es zwar schon lange durchsetzen, aber einstweilen ist mir bei Rohde noch so wohl, daſs ich diese Schule gar nicht so bald abbrechen mag. Nach meinem Examen, also wohl im Winter 79/80, werde ich dann mein Habilitationsgesuch an Ihre Fakultät richten, nicht um gleich zu lesen, sondern um nur überhaupt erst darin zu sein. Denn daſs ich an der Basler Universität bleibe, versteht sich von selbst, denn von meiner Mutter und Ihnen und Nietzsche gehe ich nicht fort, und wenn man mir goldene Berge verspräche, was man zwar nicht so laut sagen sollte, denn solch ein Entschluſs ist ein allzu sicherer Weg um Zeit seines Lebens Privatdozent zu bleiben; aber ich habe es Ihnen nun einmal gesagt, und ἃ γέγραφα γέγραφα.

Ihr Geburtstag soll mir immer gelobt sein als Vater des Tages an dem sie mir geschrieben haben. Zu wünschen weiſs ich Ihnen freilich nichts formulibares. Denn Ihre Gesundheit ist offenbar eximirt von all den zufälligen Plackereien, mit denen andere heimgesucht sind, und wenn Sie sich Schüler gewünscht

Christian August Lobeck (1781-1860) was a German classical scholar. Lobeck was born at Naumburg. After studying at the universities of Jena and Leipzig, he became Privatdozent at the University of Wit

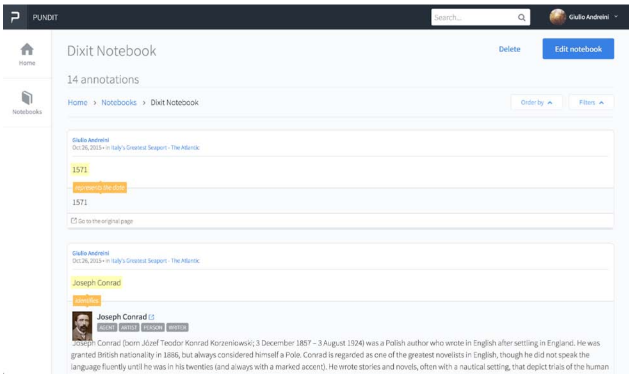
Pundit Annotator Pro used in the project The European correspondence to Jacob Burckhardt

By using Pundit, users contribute to create a new knowledge layer on top of web pages, thus enriching the web of data.

Thanks to the possibility of sharing public annotations, adding comments to or evaluating existing annotations, Pundit can be portrayed as a collaborative platform, allowing the users to cooperate and share their results with workgroups, thus enabling revisions and a form of peer reviewing. Saved annotations are collected in notebooks that can be made public or private.

Users are able to visualize and manage their annotations in the so-called 'Annotations Manager'. Besides browsing their notebooks and performing advanced research, there users can export their collected data in various standard formats. In this way, is possible to save copies of collected data, to be potentially exported in different systems, compliant with the Web Annotation standard.

The Annotation Server implements a set of REST APIs, which allow to expose public resources through content negotiation. The knowledge base can be integrated in third party projects such as Digital Libraries[1] and be used to build advanced research systems or advanced semantic data visualizations.



Joseph Conrad (born Józef Teodor Konrad Korzeniowski; 3 December 1857 – 3 August 1924) was a Polish author who wrote in English after settling in England. He was granted British nationality in 1886, but always considered himself a Pole. Conrad is regarded as one of the greatest novelists in English, though he did not speak the language fluently until he was in his twenties (and always with a marked accent). He wrote stories and novels, often with a nautical setting, that depict trials of the human

The Pundit Annotations Manager is the application where users can manage their annotations

In these last years the software underwent continuous refactoring with the objective to make it more and more modular and adaptable in several contexts. Besides the graphical interface flexibility, with its pro and light modes, the client allows the use of custom vocabularies and ontologies.

Pundit can be customized for tailor made projects where requirements lie outside its standard features. Starting from open source code new features not present in the original version are developed.

The first version of Pundit was developed from 2010 within the EU Semlib project (http://cordis.europa.eu/result/rcn/57391_en.html). Then in 2012 the development continued in the context of the EU DM2E project (http://dm2e.eu/). In 2014 the StoM project (http://www.stom-project.eu) started with the aim of bringing Pundit in the market as a software-as-a-service platform. Specific features of Pundit are also under development within EU Europeana Sounds (http://www.europeanasounds.eu/) project.

Pundit is used for semantic annotations in research projects such as the ERC AdG LOOKINGATWORDS (http://cordis.europa.eu/project/rcn/102545_en.html) and ERC AdG EUROCORR (http://www.burckhardtsource.org) funded projects. In both applications it is used by teams of researchers to semantically enrich the corpus of text of their digital library.

Bibliography

Morbidoni, C. and Piccioli, A. (2015). Pundit 2.0. *Semantic Web journal*, First published January 31, 2015: http://www.semantic-web-journal.net/system/files/swj1003.pdf.

Notes

[1] In particular, we integrated Pundit with Muruca, the open source Digital Library Framework developed by Net7 (http://muruca.org/en/).

Playing With Cultural Heritage Through Digital Gaming: The New Narrative of the ARK4 Project

Alexandra Angeletaki
alexandra.angeletaki@ub.ntnu.no
Norwegian University of Science and Technology (NTNU)

Agiatis Benardou
a.benardou@dcu.gr
Digital Curation Unit, Athena Research Centre, Greece

Nephelie Chatzidiakou
n.chatzidiakou@dcu.gr
Digital Curation Unit, Athena Research Centre, Greece

Eliza Papaki
a.papaki@dcu.gr
Digital Curation Unit, Athena Research Centre, Greece

In this poster we illustrate the impact of digital technology applications in the field of the contemporary museum and cultural heritage practice using data from workshops held at the 3D laboratory in Trondheim, Norway as well as in the Digital Curation Unit, ATHENA R.C. in Athens, Greece.

Significant transformations have taken place in the field of digital heritage due to the large extent of digitization of cultural heritage collections, the development of gaming and the application of more interactive use of cultural information on the Web. Going digital in the cultural heritage sector has created another space of interaction for users who seem to be increasingly involved in this digital landscape. Apart from that, digital applications have wore different disguises employed in different kinds of devices used effectively in the GLAM sector. Whether this wide adoption of technology suggests the wider engagement of the public with cultural heritage awaits interpretations.

Does a digital visit realize itself differently in an immersive cultural landscape where the person visiting a site or an exhibition is active in seeking knowledge? Does this innovation really transforms the relation between user and cultural object? These are some of the main questions the project ARK4 has been dealing with since 2014. Experimenting with content, appearance and user design, our aim was to create a new virtual space of dialogue between the person asking the question and the organisation holding the answer and to explore different methods of technology in disseminating knowledge from the past to a young audience. To this purpose, we have experimented with different kinds of content from botany to archaeology and with different types of technology applications, from digital games on touch screens to online questionnaires. By following user centric methods, this varied interplay of content, digital applications and audience has presented interesting findings in terms of user satisfaction, evaluation of the content and the digital means.

As the project enters its second phase, it also aims to create engagement and educational activities in the immediate community in order to engage school-children, university students and local enterprises in the seek of knowledge of the past. Each individual can be allowed in the frame of our project to deliver their own experiential perception on the story/game he/she chooses to interact with according to their own individualized level of pre-understanding and motivation. That is the visitor`s background, nationality and identity may influence and vary the outcome of the experience to be expected. Thus the visit becomes a complex process of interpretation and our inquiry might add a new dimension to the debate of creating a dialogue between European shared memory institutions and the individual visitor which can then manifest itself in experiencing diversity. Using the interaction between the participants and the objects as an observation field one allows the outcome to be varied and justified by the visitors' personal intention. Our poster will present findings of three workshops organized in collaboration with museums and schools in Norway and Greece between 2014 and 2015. In our study we analyze data on how school children interact, work and learn in the context of educational workshops, through observation, discussions, and direct surveys, interviews of the students, system log-files and performance tests. Focusing on archaeological context, the project ARK4 in its new phase will further explore user interactivity with digital technologies and gaming. The broader impact of the study contributes to the discussions on issues pertaining to educational activities from the users' perspective.

ITAF: Rewiring the Italian 'Nation' of the Army of Flanders (1567-1714)

Maurizio Arfaioli
m.arfaioli@gmail.com
The Medici Archive Project, Italy

Established in 1568 to crush a rebellion triggered by religious strife, which through the years turned into a conflict of empires, the Spanish Army of Flanders rose to the challenge, becoming Europe's largest standing army since Roman times. An army in which troops from all corners of the Spanish empire (and beyond) were called to serve, from Portugal to the Balkans, from Scotland to Sicily. Between 1568 and 1713 – when the Treaty of Münster put an end to the existence of the Spanish Netherlands – more than forty infantry regiments, dozens of companies of light cavalry, several thousands of gentlemen adventurers, sailors, artillerymen, administrators, ecclesiastics and architects from all corners of Spanish Italy 'passed to Flanders' (as the saying went) to help restore the authority of their king. And that without mentioning the numerous unaccounted-for women and children that followed them.

In the course of a century and a half of almost uninterrupted conflicts, the Italian military 'nation' of the Army of Flanders existed as a protean, complex and extended network, created through a largely unplanned and spontaneous process of social and cultural sedimentation by wave after wave of soldiers and officers that spent an important part of their active lives in the Low Countries. Still, owing the heavy historiographical stigma that accompanied for a long time the memory of 'Spanish Italy', the Italian military 'nation' in the Army of Flanders has gradually slipped into oblivion, and is nowadays largely forgotten.

ITAF (Italian Troops of the Army of Flanders), is an application tailored to 'rewire' the tangled web of military and social hierarchies which sustained the life of the Italian military 'nation' in the Low Countries through the effective integration of data extracted from a variety of archival and bibliographic sources. I intend to present at DH Krakow 2016 the trial version of this application/database, with which I am conducting a pilot study on the history of the longest-lived Italian infantry *terzo* (the Italianization of the Spanish word *tercio*) of the Army of Flanders – a unit that served the Spanish monarchy in the Low Countries from 1597 to 1713. The final version of this application is meant to be applied to the study of the entire Italian 'nation', and, with a few adaptations, its use could be extended to that of each of the other 'nations' (Spanish, Walloon, British, German etc.) of the Army of Flanders – or to the Army as a whole. **ITAF** is developed open source and relies on a NoSQL Database for data management.

Bibliography

Gonzalez de Leon, F. (2009). *The Road to Rocroi. Class, Culture and Command in the Spanish Army of Flanders, 1567-1659.* Leiden: Brill.

Parker, G. (1972). *The Army of Flanders and the Spanish Road 1567-1659.* Cambridge: Cambridge University Press

Documenting Material Culture: 3D Laser Scanning, Photogrammetry and Archaeological Ceramics from Lincoln Pottery Works

Effie F. Athanassopoulos
efa@unl.edu
University of Nebraska-Lincoln, United States of America

Aaron Pattee
acpattee30@gmail.com
University of Nebraska-Lincoln, United States of America

This poster explores the application of laser scanning and photogrammetric recording methods to archaeological ceramics excavated from a former pottery factory in Lincoln, Nebraska, USA. As 3D modeling methods continue to become more user-friendly and affordable, they offer an attractive alternative for artifact documentation, analysis and sharing of data, compared to the traditional methods of photography and profile drawings. Several studies have utilized 3D scanning for accurate data acquisition, including 2D profiling, and the calculation of attributes that are harder to measure by traditional means, e.g. volume, surface area, and symmetry. Furthermore, the 3D models can become part of digital publications and form the basis for the creation of new digital resources.

Here we are presenting 3D models created with two different methods: laser scanning and photogrammetry. We used a Next Engine 3D laser scanner, a portable scanner, suitable for small to medium-sized artifacts. The scanner generates 3D point clouds and also records texture. The texture is not high resolution, so it is not as sharp as high quality digital photographs. Multiple views are needed to create a complete model. The model requires editing (trimming, aligning, fusing) and, depending on the complexity of the object, it can take from one to two hours for a complete edited model. The 3D models created with the laser scanner are very accurate; they include any surface lines, indentations, breaks, imperfections, etc. We also use photogrammetry in order to improve upon texture and to compare the accuracy of the photogrammetric models with the 3D laser scanner models. All photographs are taken with a Nikon D3300 DSLR camera. Agisoft Photoscan Pro is used for the alignment, dense cloud, mesh, and texture for each model. Additional trimming and modifications are completed with CloudCompare (an open source 3D point cloud processing software). The combination of both methods allows for precise replication and creation of high resolution models. It also facilitates multiple types of analysis.

The case study is the Lincoln Pottery Works (LPW), in Lincoln, NE, USA, which operated from 1880 until around 1903. The inventory of the LPW centered on utilitarian, domestic wares. It produced crocks, jugs, bowls, jars, lids, flower pots, planters, architectural terra cotta and other types of ceramics. Ceramic technology developed rapidly during this period to include semi-mechanized means of forming pots, known as “jigging” or “jollying,” which allowed for mass production of vessels. LPW also exhibits the latest innovations in nineteenth-century kiln design in the form of downdraft kilns. When the LPW was founded, Lincoln was a prosperous and rapidly growing city with a population of c. 13,000 in 1880 which had increased to 55,000 in 1890 (Bleed and Schoen, 1990: 34). After the factory closed, most of the property became part of a housing development. The site was excavated by Peter Bleed in 1986-1987, and the results were published in 1993. The LPW collections are currently in the Nebraska State Historical Society.

This is a pilot project and a work in progress. The digital recording of representative types of LPW ceramics through 3D interactive models is the first step in the creation of an online exhibit and digital resource. Currently, we are documenting the most common types of ceramics produced by LPW. The LPW collection is extremely large. For example, the estimated number of bowls is 5,633 vessels and represent 38.45% of the assemblage. Thus, our goal is to document a representative sample of the collection, we do not plan to document the collection fully. Once the pilot phase is completed, we will make the 3D models, along with other content, available on a website (in collaboration with the Nebraska State Historical Society).

The digitization project will facilitate different kinds of analysis that build on the original publication of the results; for example, the distribution of LPW products in other Midwestern states (the records of the business have not been preserved), socioeconomic aspects, food ways, a gendered perspective, etc. The first step, however, is to create a digital resource that will draw attention to this collection and engage other researchers and the public. By utilizing the latest technology to create and present accurate models of LPW representative products, we hope to bring wider attention to this important assemblage, which is an integral part of the history of nineteenth century Nebraska and the industrial archaeology of the Great Plains region.

Bibliography

Bleed, P. and Schoen, Ch. (1990). The Lincoln Pottery Works: A Historical Perspective. *Nebraska History*, **71**: 34-44.

Schoen, Ch. and Bleed, P. (1993). The Archaeology of the Lincoln Pottery Works. *Central Plains Archaeology*, **3**(1): 1-240.

Extraction and Visualization of Toponyms in Diachronic Text Corpora

Adrien Barbaresi
adrien.barbaresi@oeaw.ac.at
Austrian Academy of Sciences, Austria; Berlin-Brandenburg Academy of Sciences, Germany

Hanno Biber
hanno.biber@oeaw.ac.at
Austrian Academy of Sciences, Austria

Introduction

This paper focuses on the extraction of German and Austrian place names in historical texts. It is part of a cooperation between the Berlin-Brandenburg and the Austrian Academies of Sciences. The latter is the holder of the text basis for this investigation, the digitized version of the satirical literary magazine "Die Fackel" ("The Torch"). It has been originally published and almost entirely written by the satirist and language critic Karl Kraus in Vienna from 1899 until 1936, and contains a considerable variety of toponyms (Biber, 2001).

Gazetteers and lists

Toponym resolution often relies on named-entity recognition and artificial intelligence (Leidner and Lieberman, 2011). However, knowledge-based methods using fine-grained data – for example from Wikipedia – have already been used with encouraging results (Hu et al., 2014).During the 20th century there have been significant political changes in Central Europe that have severely affected toponyms, so that geographical databases lack coverage and detail. Consequently, the database we develop follows from a combination of approaches: gazetteers are curated in a supervised way to account for historical differences,and current geographical information is used as a fallback.

First, a cascade of filtersis used: (1) current and historical states (e.g. Austria-Hungary); (2) regions, important subparts of states, and regional landscapes (e.g. Swabia); (3) populated places; (4) geographical features (e.g. valleys). Wikipedia's API is used to navigate in categories and to retrieve coordinates, which are completed by hand for states and regions.Second, current information is also compiled from the Geonames database[1]: data for European countries are retrieved and preprocessed(variants and place types).

In order to exclude common and proper nouns, the German version of the Wiktionary serves as a reference[2], and registers of frequent surnamesand family names, as well as well-known persons (especially writers) are built using Wikipedia and Wikidata.

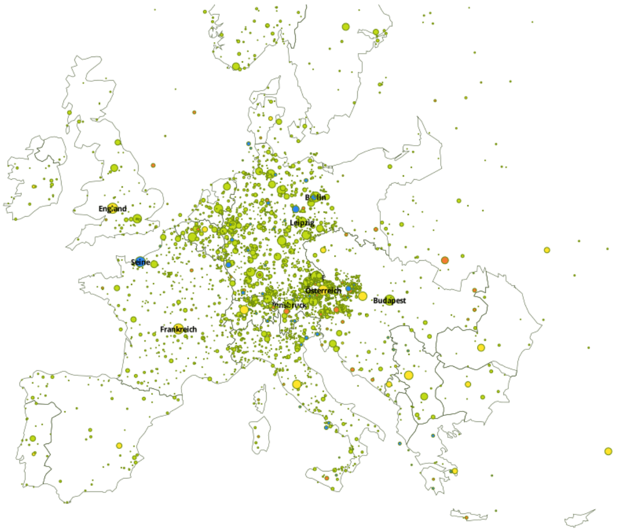
Extraction

The texts were digitized, manually corrected as well as manually annotated with respect to the names of persons and institutions, so that most proper nouns which are not place names can be excluded from the study.

The tokenized files of works to be analyzed (Jurish and Würzner, 2013) are filtered and matched with the databaseby finite-state automatons: toponyms are extracted using a sliding window (for multi-word names up to three components). Disambiguation being a critical component (Leetaru, 2012), an algorithm similar to Pouliquen et al. (2006), who demonstrated that an acceptable precision can be reached that way,guesses the most probable entry based on distanceto Vienna (Sinnott, 1984), contextual information(closest-country, last names resolved), and importance (place type, population count).

Visualization

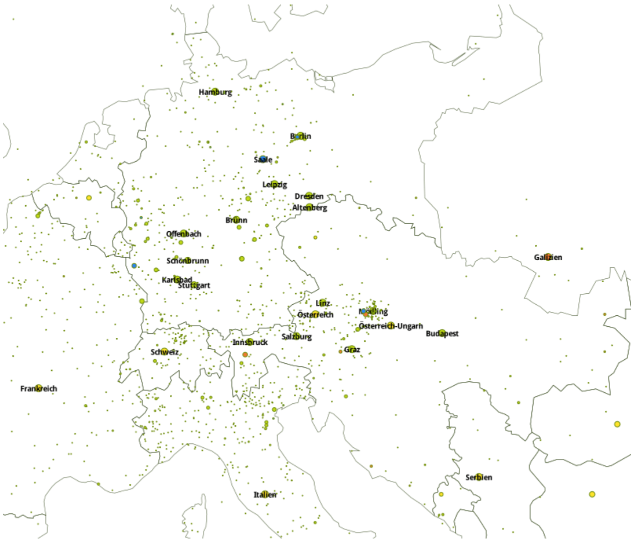
The results are projected on a map of Europe with boundaries of 1914[3] using TileMill[4]. They are customized with CartoCSS: multiple trial-and-error iterations are performed concerning both data quality and graphical output. The two experimental maps belowground on the same data, they result from different settings.



Map 1– Experiment on European scale with boundaries of 1914 (yellow: sovereign territories; orange: regions; green: populated places; blue: geographical features)

Discussion

Potential conceptual caveats include previous times as well as fictitious places, especially names which can refer to mythological and actual places of Ancient Greece or Rome. Practical caveats are for instance false localizations due to disambiguation errors (e.g. Brünn/Brno on map 2). We plan to bypass the disambiguation for a hand-picked list of places. As big datais an entanglement of implicit theoretical assumptions (Crawford et al., 2014), the difference between a mere data collection project and a digital humanities study resides precisely in the number and diversity of filters used. The code and listings produced for this study are available online.[5] We plan to integrate corpora of greater variety and scope and to include more specific metadata in order to design versatile visualizations.



Map 2– Central Europe, experiment with a restrictive filtering

Conclusion

A map is a discrimination, a marking of difference (Wulfman, 2014): our maps highlight the linguistic and cultural ties of Kraus and his contemporaries with Bohemia and Northern Italy, where there are more numerous toponyms to be found than in Hungary.Beyond that, "Die Fackel" is (at least) a European phenomenon; Kraus' vision of Europe is more inclined towards cultural centers (Prague, Munich, Paris, Berlin).

It is our hope that visualization studies based upon mixed methods contribute to a greater awareness of the potential of digital heritage as well as literary studies in the digital age. Although the maps seem immediately interpretable, they are not an objective result but a construct (Juvan, 2015), the result of a filtering. The "human" interventions on the map as well as the technical competence to do so replace this study in the hermeneutic circle of the philological tradition.

Bibliography

AAC-FACKEL. Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936. In Biber, H., Breiteneder, E., Kabas, H., Mörth, K.; Graphic Design: Burdick, A.(eds), *AAC Digital Edition 1*, http://www.aac.ac.at/fackel.

Biber, H. (2001). In Wien, in Prag und infolgedessen in Berlin - Ortskonstellationen in der "Fackel". In Marten-Finnis, S., Uecker, M. (ed.) *Berlin-Wien-Prag. Moderne, Minderheiten und Migration in der Zwischenkriegszeit*, Peter Lang, pp. 15-26.

Crawford, K., Gray, M. and Miltner, K. (2014). Big Data | Critiquing Big Data: Politics, Ethics, Epistemology | Special Section Introduction. *International Journal of Communication*, **8**: 1663–72.

Hu, Y., Janowicz, K. and Prasad, S. (2014). Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia. *Proceedings of the 8th Workshop on Geographic Information Retrieval*, ACM, pp. 8-16.

Juvan, M. (2015). From Spatial Turn to GIS-Mapping of Literary Cultures. *European Review* **23**(1): 81-96.

Jurish, B. and Würzner, K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, **28**(2): 61–83.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, **3**(2): 5-11.

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fluart, F., Zaghouani, W., Widiger, A., Forslund, A.-C. and Best, C. (2006). Geocoding multilingual texts: Recognition, disambiguation and visualisation. *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 53-58.

Sinnott, R. W. (1984). Virtues of the Haversine. *Sky and Telescope* **68**(2): 159.

Wulfman, C. E. (2014). The Plot of the Plot: Graphs and Visualizations. *The Journal of Modern Periodical Studies*, **5**(1): 94-109.

Notes

[1] http://www.geonames.org/

[2] Thanks to Kay-Michael Würzner (BBAW) for his extraction script.

[3] http://dev2.dariah.eu/geoserver/web/

[4] https://www.mapbox.com/tilemill/

[5] https://github.com/adbar/toponyms

Named Entity Extraction from digitized texts of Mongolian Historical Documents in Traditional Mongolian Script

Biligsaikhan Batjargal
biligsaikhan@gmail.com
Research Organization of Science and Engineering, Ritsumeikan University, Japan

Garmaabazar Khaltarkhuu
garmaabazar@gmail.com
Mongolia-Japan Center for Human Resources Development, Mongolia

Akira Maeda
amaeda@media.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

In this paper, we demonstrate a named entity extraction method for digitized ancient Mongolian documents by using features of traditional Mongolian script. In the field of humanities, getting knowledge by analyzing various historical documents is an important task. There are increasing demands from Mongolian humanities researchers to perform text analysis at massive scale with prompt and accurate results. A few ancient Mongolian historical manuscripts including 1) the "Qad-un ündüsün-ü quriyangγui altan tobči neretü sudur (The Altan Tobchi or the Golden Summary: Short history of the Origins of the Khans)" a.k.a "Little" Altan Tobchi, and 2) the "Asaraġci neretü-yin teüke or Asragch nėrtiin tüükh (The Story of Asragch)", which were written in traditional Mongolian script have been converted to digital texts and made publicly available through the traditional Mongolian script digital library (TMSDL) (Batjargal et al., 2013). Figure 1 shows a page of the "Little" Altan Tobchi in the TMSDL. The demands from Mongolian humanities researchers, as well as the lessons learned from the TMSDL have encouraged us to conduct further research in developing a new method for extracting named entities from ancient Mongolian historical documents. However, there has been little research on text mining or named entity extraction for Mongolian language and none of the research has considered text mining on ancient Mongolian historical documents due to the lack of research in those areas. Thus, we want to propose a named entity extraction method for ancient historical documents in traditional Mongolian script by employing machine learning techniques for aiming to reduce the labor-intensive analysis on historical text.

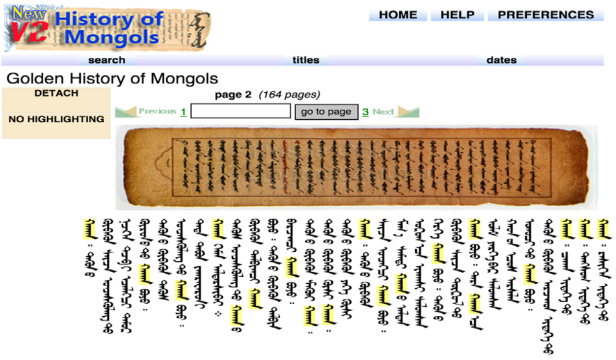


Figure 1. Screenshot of the TMSDL

In the proposed approach, an ancient Mongolian corpus gets tokenized, each token gets annotated and gold standard annotations are prepared for inputting into computer system for learning. The proposed method learns the extraction rules of personal names and place names from annotated training corpora, and then extracts named entities from ancient Mongolian texts by employing machine learning techniques (Batjargal et al., 2015).

We use the IOB2 (Ramshaw and Marcus, 1995) format for tagging tokens. Because of some unique features of traditional Mongolian script, we also use "Start/End" (SE) chunk tag set (Asahara and Matsumoto, 2003). "S" tag is attached to the first character of each word including the named entities and "E" tag to the last character. Thus, each token will include the 1) IOB2 tag and 2) SE tag.

We also consider the following features of traditional Mongolian script for differentiating personal names and place names.

- **Information of the preceding** and **following tokens:** Features are extracted by looking the context of the current, preceding, and succeeding tokens. If the preceding token is generational or dynastic information, an inherited or life-time title of nobility, or a traditional descriptive phrase, it could indicate the current token is a personal name.
- **Suffix:** Many living being and humankind proper names take only certain plural suffixes such as ᠨᠠᠷ (nar/ner) and possessive suffixes (Chinggaltai, 1963). Some suffixes are visually separated from the stem of a word or other suffixes, but any attached suffixes are considered to be an integral part of the word.
- **Beginning of a sentence:** Subjects or personal names often appear at the beginning of a sentence.
- **End of a token:** Words with a final vowel letter 'a' or 'e' are separated visually from the preceding consonant by a narrow gap. However, the 'a' or 'e' is an integral part of the word stem.

For evaluation, we calculated precision, recall, and F-measure by the 5-fold cross-validation. To prepare the gold standard annotations, we annotated all the personal names and place names in the "Little" Altan Tobchi using the manually compiled personal and place names' indices obtained from the "Qad-un ündüsün quriyangγui altan tobči –Textological Study" (Choimaa, 2002). For the experimental corpus, we utilized digitized text of chronological manuscripts "Little" Altan Tobchi. We utilized the LIBLINEAR with the L2-regularized L2-loss support vector classification (dual) solver (Rong-En Fan et al., 2008).

We will further improve the proposed method by considering more features by conducting various experiments with different combinations of features for checking whether the particular feature set will improve the preliminary results of 0.70 of precision, 0.57 of recall and 0.63 of F-measure or not.

Bibliography

Asahara, M. and Matsumoto, Y. (2003). Japanese Named Entity Extraction with Redundant Morphological Analysis, *Proceeding of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*, Stroudsburg, PA, USA, June 2003, pp. 8–15.

Batjargal, B., Khaltarkhuu, G., Kimura, F. and Maeda, A. (2012). Developing a Digital Library of Historical Records in Traditional Mongolian Script, *International Journal of Digital Library Systems*, **3**(1): 33–53.

Batjargal, B., Khaltarkhuu, G., Kimura, F. and Maeda, A. (2015). An Approach to Named Entity Extraction from Mongolian Historical Documents, *Proceedings of the International Conference on Culture and Computing (Culture and Computing 2015)*, Kyoto, Japan, October 2015, pp. 205-06.

Chinggaltai. (1963). *A Grammar of the Mongol Language.* New York: Frederick Ungar Publishing Co.

Choimaa, Sh. (2002). *Qad-un űndűsűn quriyangyui altan tobči (Textological Study)* vol. **1**. (in Mongolian). Ulaanbaatar: Centre for Mongol Studies, National University of Mongolia, Urlakh Erdem.

Fan, R.-E., Chang, K.–W., Hsieh, C.–J., Wang, X.–R. and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, **9**: 1871–74.

Ramshaw, L. A. and Marcus, M. P. (1995). Text Chunking Using Transformation-based Learning, *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, Cambridge MA, USA, June 1995, pp. 82–94.

Interoperability: a new horizon for data sharing in Humanities and Social Sciences. The input of three digital services developed by Huma-Num

Olivier Baude
Olivier.Baude@huma-num.fr
TGIR Huma-Num UMS 3598 CNRS (Centre National de la Recherche Scientifique), France

Adeline Joffres
adeline.joffres@huma-num.fr
TGIR Huma-Num UMS 3598 CNRS (Centre National de la Recherche Scientifique), France

Nicolas Larrousse
nicolas.larrousse@huma-num.fr
TGIR Huma-Num UMS 3598 CNRS (Centre National de la Recherche Scientifique), France

Stéphane Pouyllau
stephane.pouyllau@huma-num.fr
TGIR Huma-Num UMS 3598 CNRS (Centre National de la Recherche Scientifique), France

In the field of Humanities and Social Sciences, the production of digital or scanned data has increased considerably in recent years. These data, which are usually very expensive to produce, are often lost at the end of the project. They are therefore rarely reused, due to a lack of financial, human and technical resources of the communities that produced them.

The TGIR Huma-Num, whose mission is to facilitate the digital turn in Humanities and Social Sciences research, offers services dedicated to the production and reuse of data. These services aim at avoiding loss and facilitating the re-use of scientific data. To do this, Huma-Num supports research teams throughout their digital projects to allow the sharing, reuse and preservation of data thanks to a chain of devices focused on interoperability.

The aim is to foster the exchange and dissemination of metadata, but also of data itself via standardized tools and lasting, open formats. These tools developed by Huma-Num are all based on Semantic Web technologies, mainly for their auto-descriptive features and for the enrichment opportunities they enable. Other interoperability technologies complement them, such as the OAI-PMH. Interoperability is used internally to allow Huma-Num services to communicate with one another and externally to let users plug their tools into Huma-Num services.

Another important point is to make the storage of data independent of the device used to disseminate the data.

This poster will present three services designed and developed by Huma-Num to process, store and display research data while preparing them for re-use and long-term preservation.

These services embrace the research data life cycle and are designed to meet the needs arising therefrom:

- SHARE data and metadata, using interoperable technologies, with NAKALA. Another feature is the possibility to make DATA CITEABLE with PIDs;
- SHOW and DISPLAY the data with NAKALONA, using the CMS Omeka allowing customised editorialization of data stored in Nakala (e.g. virtual exhibitions) benefiting from the features of this CMS such as its search engine and extended OAI-PMH feeds which facilitate interoperability;
- TAG and PUSH data through ISIDORE, enriching and interconnecting them to ensure better visibility.

These three complementary services thus constitute a coherent chain of research data tools. While they interact smoothly with each other, they are also open to external tools using the same technologies. In fact each tool, considered individually, is not ground-breaking, but we consider that the combination of these tools is the key to address needs and to prepare the long term preservation of scientific data.

The scientific objective is to promote data sharing so that other researchers, communities, or disciplines, can reuse them, including from an interdisciplinary perspective and in different ways. A map, for example, may become a scientific object that reflects both the point of view of a geographer and that of a historian. More generally, the principles and methods of the Semantic Web (RDF, SPARQL, SKOS, OWL) on which these services rely enable data to be documented or re-documented for various uses without confining them to inaccessible silos.

The second objective is to prevent the loss of data by preparing their long-term preservation. Documenting the use of appropriate formats, which are the basis of data interoperability, greatly facilitates the archiving process.

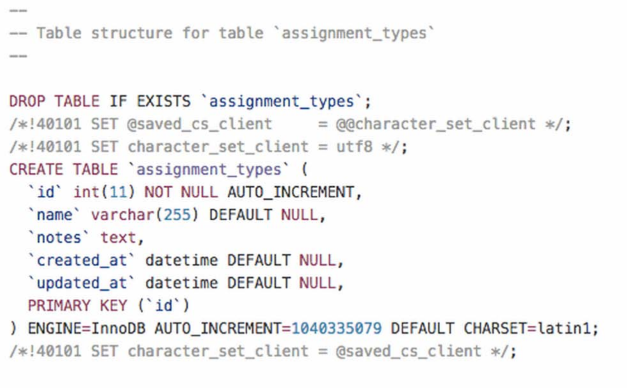
Through these services, the TGIR Huma-Num is developing solutions based on digital technologies and, in particular, those of the Semantic Web, to meet the needs of researchers and scientific communities, and make new research in the field of Humanities and Social Sciences possible.

Seeing the Argument: Visualize Your Database with DAVILA

Jean Ann Bauer
jabauer@princeton.edu
Princeton University, United States of America

Data modeling is a subject of study within digital humanities (Flanders and Jannidis, 2015) with special emphasis on the creative, scholarly, and normative aspects of data modeling in database design (Verhoeven, 2014; Bauer, 2015). By imagining our sources as data, and then abstracting th ose data to the point where quantitative methods gain purchase but still retain their original context, digital humanists are reimagining what it means to study the human record. While the study of data modeling is crucial to digital humanities, it is complicated by database schema conventions. Whether expressed in technical diagrams or computer code, database schemas initially present as confusing to anyone not trained to read them. Yet, schemas contain crucial information about how humanities data are translated to the computer, creating affordances for querying, browsing, and display – making direct access to schemas valuable to researchers, users, and peer reviewers.

DAVILA is an open source relational database schema annotation and visualization tool created to help bridge this gap. While there are other database schema visualization tools (ex. MySQL Workbench, DbVisualizer, and DbSchema) available, DAVILA was specificallydesigned tofunction within digital humanities project teams and to document digital humanities projects for the non-technical audiences that use and review them. Drawing onthe scholarly practice of annotation as well as the principles of Information Design (Tufte, 2006a; Tufte, 1997; Tufte, 2006b; Tufte, 2006c) and Computational Information Design (Fry, 2004), DAVILA creates a modified Entity-Relationship diagram (see (Ramsay, 2004) for a more detailed description of E-R diagrams) that can be used by both technical and content experts.



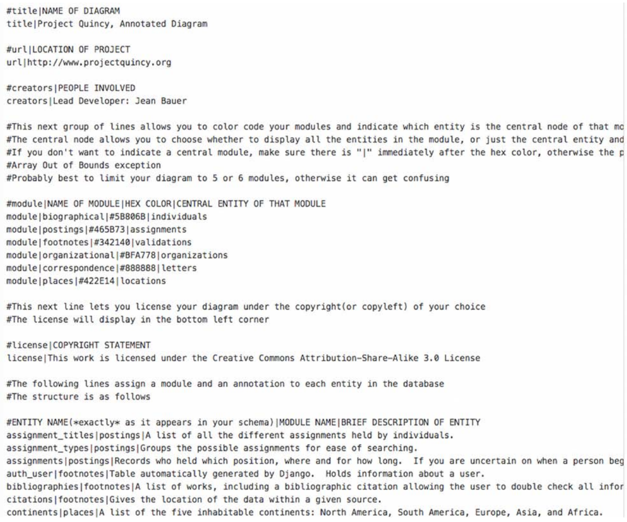
```
--
-- Table structure for table `assignment_types`
--

DROP TABLE IF EXISTS `assignment_types`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `assignment_types` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `name` varchar(255) DEFAULT NULL,
  `notes` text,
  `created_at` datetime DEFAULT NULL,
  `updated_at` datetime DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB AUTO_INCREMENT=1040335079 DEFAULT CHARSET=latin1;
/*!40101 SET character_set_client = @saved_cs_client */;
```

Figure 1: A sample database schema, in this case the output from a mysqldump file for Project Quincy (http://projectquincy.org)

DAVILA takes two files as inputs: a database schema and an optional plain text customization file. Examples of both files are included with the software. When a schema is processed without a customization file DAVILA will create a blue box for each entity, and connect the entities based on their relationships (if specified in the schema). So, if a user is given someone else's database without documentation, she can use DAVILA to create a technical diagram from the schema. The user can then explore the entities in the interface and get a better sense of a new system. The customization file can be completed in an iterative process until the entire system has been documented.

The customization file is the heart of DAVILA, and what separates it from other database schema visualization software packages. The plain text file allows scholars to annotate the resulting diagram. These annotations begin with basic project metadata, crucial to data curation (Flanders and Muñoz, 2014): project name, URL, developer names, and the copyright license applied to the diagrams. The customization file allows users to group entities into modules, color code those modules, indicate which entity is central to each module, and provide annotation text for every entity in the database. The annotated text is particularly crucial, describing in a sentence or two the essence of a particular entity and its relationship to the larger database. The color coding and positioning bring order to the diagram, but the annotation gives non-technical team members or peer reviewers a way into the schema.

Once DAVILA is running, users can click and drag the entities into different positions, expand an individual module for more information, or hide the non-central entities in a module to focus on another part of your schema, all in a fun, force-directed environment courtesy of the toxiclibs physics library. Pressing the space bar saves a snapshot of the window as a timestamped, vector-scaled pdf. As a visualization tool, DAVILA has limited use for a team that combines sighted and blind (or colorblind) team members. For such groups, the customization file may provide a more fruitful locus of collaboration than the visual output.



```
#title|NAME OF DIAGRAM
title|Project Quincy, Annotated Diagram

#url|LOCATION OF PROJECT
url|http://www.projectquincy.org

#creators|PEOPLE INVOLVED
creators|Lead Developer: Jean Bauer

#This next group of lines allows you to color code your modules and indicate which entity is the central node of that mo
#The central node allows you to choose whether to display all the entities in the module, or just the central entity and
#If you don't want to indicate a central module, make sure there is "|" immediately after the hex color, otherwise the p
#Array Out of Bounds exception
#Probably best to limit your diagram to 5 or 6 modules, otherwise it can get confusing

#module|NAME OF MODULE|HEX COLOR|CENTRAL ENTITY OF THAT MODULE
module|biographical|#5B806B|individuals
module|postings|#465B73|assignments
module|footnotes|#342140|validations
module|organizational|#BFA778|organizations
module|correspondence|#888888|letters
module|places|#422E14|locations

#This next line lets you license your diagram under the copyright(or copyleft) of your choice
#The license will display in the bottom left corner

#license|COPYRIGHT STATEMENT
license|This work is licensed under the Creative Commons Attribution-Share-Alike 3.0 License

#The following lines assign a module and an annotation to each entity in the database
#The structure is as follows

#ENTITY NAME(*exactly* as it appears in your schema)|MODULE NAME|BRIEF DESCRIPTION OF ENTITY
assignment_titles|postings|A list of all the different assignments held by individuals.
assignment_types|postings|Groups the possible assignments for ease of searching.
assignments|postings|Records who held which position, where and for how long.  If you are uncertain on when a person beg
auth_user|footnotes|Table automatically generated by Django.  Holds information about a user.
bibliographies|footnotes|A list of works, including a bibliographic citation allowing the user to double check all infor
citations|footnotes|Gives the location of the data within a given source.
continents|places|A list of the five inhabitable continents: North America, South America, Europe, Asia, and Africa.
```

Figure 2: Sample customization file that ships with DAVILA. Note that the diagram now has associated metadata, including a copyright statement

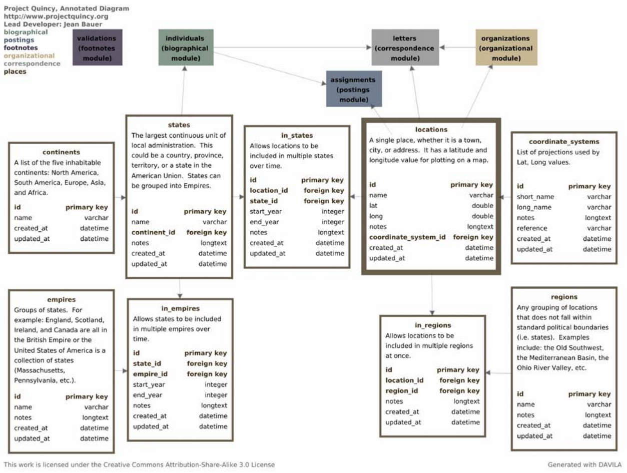

Project Quincy, Annotated Diagram
http://www.projectquincy.org
Lead Developer: Jean Bauer
biographical
postings
footnotes
organizational
correspondence
places
validations (footnotes module)
individuals (biographical module)
letters (correspondence module)
organizations (organizational module)
assignments (postings module)
states
The largest continuous unit of local administration. This could be a country, province, territory, or a state in the American Union. States can be grouped into Empires.
id primary key
name varchar
continent_id foreign key
notes longtext
created_at datetime
updated_at datetime
continents
A list of the five inhabitable continents: North America, South America, Europe, Asia, and Africa.
id primary key
name varchar
created_at datetime
updated_at datetime
in_states
Allows locations to be included in multiple states over time.
id primary key
location_id foreign key
state_id foreign key
start_year integer
end_year integer
notes longtext
created_at datetime
updated_at datetime
locations
A single place, whether it is a town, city, or address. It has a latitude and longitude value for plotting on a map.
id primary key
name varchar
lat double
long double
notes longtext
coordinate_system_id foreign key
created_at datetime
updated_at datetime
coordinate_systems
List of projections used by Lat, Long values.
id primary key
short_name varchar
long_name varchar
notes longtext
reference varchar
created_at datetime
updated_at datetime
empires
Groups of states. For example: England, Scotland, Ireland, and Canada are all in the British Empire or the United States of America is a collection of states (Massachusetts, Pennsylvania, etc.).
id primary key
name varchar
notes longtext
created_at datetime
updated_at datetime
in_empires
Allows states to be included in multiple empires over time.
id primary key
state_id foreign key
empire_id foreign key
start_year integer
end_year integer
notes longtext
created_at datetime
updated_at datetime
in_regions
Allows locations to be included in multiple regions at once.
id primary key
location_id foreign key
region_id foreign key
notes longtext
created_at datetime
updated_at datetime
regions
Any grouping of locations that does not fall within standard political boundaries (i.e. states). Examples include: the Old Southwest, the Mediterranean Basin, the Ohio River Valley, etc.
id primary key
name varchar
notes longtext
created_at datetime
updated_at datetime
This work is licensed under the Creative Commons Attribution-Share-Alike 3.0 License
Generated with DAVILA

Figure 3: Example of the "Locations Module" in Project Quincy laid out in DAVILA. Note that the other modules are still in the visualization, but have been minimized. Also note that primary keys are highlighted, relationships have directionality, and datatypes are specified for each attribute.

DAVILA is written in Processing, a Java-based artistic scripting language originally created by Ben Fry and Casey Reas, and released under GPLv3. Special attention was paid to documenting DAVILA itself. Each file in the program starts with at least one paragraph, in English, describing what the following code does and how to modify it as needed. This supplements the extensive inline commenting of the code, which can double as a tutorial for >Processing. Finally, there is a detailed README file which walks new users through installing Processing and using DAVILA with their own database schemas.

DAVILA-generated schemas are used to teach database design and were included in a database inflected doctoral dissertation in history (Bauer, 2015). To generate your own diagram visit.

Bibliography

Bauer, J. (2015). Republicans of Letters: The Early American Foreign Service as Information Network, 1775-1825 University of Virginia Ph.D. http://libra.virginia.edu/catalog/libra-oa:9454.

Flanders, J. and Jannidis, F. (2015). *Knowledge Organization and Data Modeling in the Humanities*. http://www.wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf.

Flanders, J. and Muñoz, T. (2014). An Introduction to Humanities Data Curation. *DH Curation Guide* http://guide.dhcuration.org/contents/intro/ (accessed 6 March 2016).

Fry, B. J. (2004). *Computational information design*. Massachusetts Institute of Technology.

Ramsay, S. (2004). Databases. *A Companion to Digital Humanities*. Oxford: Blackwell.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn.: Graphics Press.

Tufte, E. R. (2006a). *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, Conn.: Graphics Press.

Tufte, E. R. (2006b). *Beautiful Evidence*. Cheshire, Conn.: Graphics Press.

Tufte, E. R. (2006c). *Envisioning Information*. Cheshire, Conn.: Graphics Press.

Verhoeven, D. (2014). Doing the sheep good: facilitating engagement in digital humanities and creative arts research. *Advancing Digital Humanities: Research, Methods, Theories*. pp. 206–20.

First We Feel Then We Fall – Multimedia Adaptation of Joyce's Finnegans Wake

Katarzyna Maria Bazarnik
k.bazarnik@uj.edu.pl
Jagiellonian University, Poland

Jakub Wróblewski
jakub.a.wroblewski@gmail.com
Academy of Fine Arts in Warsaw, Poland

Finnegans Wake is the last, most mysterious book by the Irish writer James Joyce. Usually described as a novel, it is a fascinating text written primarily in English (more strictly Hiberno-English), but the words are often fused with any of several dozen languages (McHugh, 1991). This dream-like narrative features a Dublin pub owner Humphrey Chimpden Earwicker, his wife Anna Livia Plurabelle, their twin sons Shem and Shaun and their daughter Issy. They travel through space and time to discover the truth about a scandalous incident in Phoenix Park in which HCE was implicated and to deliver a letter written by ALP in his defense. Drawn into a whirlpool of the past, they metamorphose into historical and legendary figures, a hill, a river, a cloud, a tree and a stone (Campbell and Robinson, 1944/1959; Bishop, 1986). The story of HCE's fall overlaps with the story of a drunken bricklayer Tim Finnegan who fell off a ladder but was resurrected when whisky splashed on his face during a fight at his wake. This fuses with the original Fall, with sexual falls of politicians and celebrities, with Napoleon and Wellington on the battlefield of Waterloo, with Tristan and Iseult's romance, a hen discovering an ancient manuscript in a rubbish heap, and more. Full of sexual innuendoes, historical, literary, autobiographical allusions and hilarious wordplays, multiple plots of *Finnegans Wake* develop in non-linear ways and can be followed like a maze, or a hypertext (Hart, 1962; Hayman, 1978; Loska, 1999 and 2000; Armand, 2003; Bazarnik, 2011).

In *The Middle Ages of James Joyce. The Aesthetics of Chaosmos* Umberto Eco offers a diagram that visualizes hypertextual complexity of the Wakean language (Eco, 1989). It shows how MEANDERTALE and MEANDERTHALLTALE, two of innumerable puns making up the textual labyrinth of *Finnegans Wake*, can be unpacked into separate words. The image represents a network connecting major components: *meander*, German *Tal*, *tale*, and *Neandertal*, which combine to suggest a cluster of meanings. This Wakean pun nudges us about how to read the book: as a "tall tale" wandering waywardly, looping backward and flashing forward, into the pre-historic past, and the origins of the human species. And as a watercourse – of course, 'the Meander' is a river which (giving us the word) meanders. Working on the logic of associations, it hints at different interpretative paths, visualised by the wavy lines of the diagram.

In our project – *First We Feel Then We Fall* – we aspire to offer a similar, dynamic, visual translation of hypertextuality and simultaneity of *Finnegans Wake* (Joyce, 1939/2002). Inspired by Eco's diagram, our project is based on a comparable analysis of narrative streams in Joyce's text. Having analysed the Wakean imagery, euphonies, rhythms and polyphonic contexts, we have selected four narrative strands, or "plots," which are translated into an up-to-date audiovisual form of an interactive Internet app. Thus networks of linguistic, historical, symbolic, and mathematical meanings entailed in Wakean puns are transposed into a dynamic audio-visual structure that the audience can co-shape in the process of interactive viewing. The viewer can switch at will between four simultaneous streams of film clips accompanied by sound (and optional captions with the *FW* text). The interactive and immersive nature of *First We Feel Then We Fall* will go beyond previous cinematic adaptations of Joyce's novel. It is the advance of digital technologies that has enabled us to approach complexity of *Finnegans Wake* in this novel way.

The application devised for this project is a multichannel, interactive video app. It will enable the viewers to experience Joyce's text in an audiovisual format consisting of simultaneously running streams, and thanks to this, offer them a portmanteaux-like audio-visual experience. The app will also contain an option suggesting a narrative path modelled on selections made by previous viewers. It will be available on the Internet browsers and mobile devices (Wróblewski, 2015). The software used includes: Final Cut for editing; Adobe After Effects and Processing for animation, and DaVinci Resolve for colour grading. Video capture was made with Canon 5D Mark III, Sony A7SII, Sony FX7e, Phantom HD Gold, and Found Footage, whereas the website was developed with Python, Django, JavaScript, HTML 5, and CSS3.

In our presentation we will briefly describe the rationale for our multimedia adaptation of *Finnegans Wake*, and let the audience experience it in an individual, interactive viewing on an available mobile device and/or a computer.

Bibliography

Armand, L. (2003). *Technē. James Joyce, Hypertext and Technology.* Prague: Karolinum.

Bazarnik, K. (2007). Joyce, Liberature and Writing of the Book. *Hypermedia Joyce Studies*, **8**(2). http://hjs.ff.cuni.cz/archives/v8_2/main/essays.php?essay=bazarnik (accessed 4 March 2016).

Bazarnik, K. (2011). *Joyce and Liberature*. Prague: Litteraria Pragensia.

Bishop, J. (1986). *Joyce's Book of the Dark, FINNEGANS WAKE.* Madison, Wisc.: The University of Wisconsin Press.

Campbell, J. and Robinson, H. M. (1944/1959). *A Skeleton Key to Finnegans Wake*. London: Faber and Faber.

Eco, U. (1989). *The Aesthetics of Chaosmos: The Middle Ages of James Joyce*. Trans. E. Esrock and D. Robey. London: Hutchinson Radius.

Hart, C. (1962). *Structure and Motif in FINNEGANS WAKE.* London: Faber.

Hayman, D. (1978). Nodality and the Infra-Structure of FINNEGANS WAKE. *James Joyce Quarterly*, **16**(1/2): 135-50.

Joyce, J. (1939/1989). *Finnegans Wake*. London: and Faber.

Joyce, J. (1939/2002). Finnegans Wake. Theall J. B. and Theall D. F. (eds).A Webified version of James Joyce's *Finnegans Wake*. Sd2cx1.webring.org. http://sd2cx1.webring.org/l/rd?ring=jamesjoyce;id=4;url=http%3A%2F%2Fwww.trentu.ca%2Ffaculty%2Fjjoyce%2Ffw.htm (accesed 18 June 2015).

Loska, K. (1999). *Wokół FINNEGANS WAKE. James Joyce i komunikacja wizualna*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Loska, K. (2000). *Finnegans Wake: James Joyce a rozumienie i interpretacja*. Kraków: Rabid.

McHugh, R. (1991). *Annotations to* Finnegans Wake. Revised ed. Baltimore: John Hopkins UP.

Wróblewski, J. (2015). *First We Feel Then We Fall teaser #1.*
Adaptation of James Joyce's Finnegans Wake by J. Wróblewski and K. Bazarnik.[video] https://vimeo.com/137422246 (accessed 31 Oct 2015).

Boutique Big Data: Reintegrating Close and Distant Reading of 19th-Century Newspapers

M. H. Beals
m.h.beals@lboro.ac.uk
Loughborough University, United Kingdom

From their earliest incarnations in the seventeenth-century, through their Georgian expansion into provincial and colonial markets and culminating in their late-Victorian transformation into New Journalism, British newspapers have relied upon scissors-and-paste journalism to meet consumer demands for the latest political intelligence and diverting content. Although this practice, wherein one newspaper extracted or wholly duplicated content from another, is well known to scholars of the periodical press, in-depth analysis of the process is hindered by the lack of formal records relating to the reprinting process. Although anecdotes abound, attributions were rarely and inconsistently given and, with no legal requirement to recompense the original author, formal records of where material was obtained were unnecessary. Even if they had existed, the number of titles that relied upon reprinted material makes systematic analysis impossible; for many periodicals, only a few issues, let alone business records, survive. However, mass digitisation of these periodicals, in both photographic and machine-readable form, offers historians a new opportunity to rediscover the mechanics of nineteenth-century reprinting. By undertaking multimodal and multi-scalar analyses of digitised periodicals, we can begin to reconstruct the precise journeys these texts took from their first appearance to their multiple ends.

Before the advent of the telegraph, individual texts were disseminated manually, through postal and private correspondence routes, over sea and land. This allowed for the relatively slow spread of texts across communication networks, as well their adaptation, truncation and expansion various stages. In a manner similar to modern internet memes, blogs and online news content, texts underwent evolutionary changes with each reprinting. These could be minute, such as the correction of spelling errors or the application of house style, or significant, through selective reordering and truncation to alter the overall meaning of the text. While identifying meme families, or collections of related texts, can help us understand what made particularly texts popular, or viral, it is only by tracing the specific trajectories and pathways of these texts that the causes and consequences of evolutionary changes can be understood.

Doing so requires us to approach these texts on multiple scales. First, by mining extremely large corpora, derived from several independent collections, we are able to identify a statistically sufficient portion of the historical network. Then, by carefully analysing the chronology and discrepancies between these reprints, hypotheses regarding institutional and industry standards can be posited and tested against the wider corpus. These efforts can be further buttressed by utilising manual transcriptions found in the personal archives of researchers using historical newspapers, such as the Scissors and Paste Database (www.scissorsandpaste.net). These transcriptions, far more accurate than the majority of datasets derived from optical character recognition, greatly improve the mining of the corpora, yielding a more complete initial network to analyse, as well as offset the skewing effect of the 'offline penumbra'.

This poster will explore the possibilities of large-scale reprint identification within and across digitised collections using a combination of Lou Bloomfield's *Copyfind* and project-specific code to identify matches between individual articles or full pages of texts in both manual (perfect) and OCR (messy) transcriptions. Exemplar collections include the British Library's 19th-Century Newspapers digital collection and planned expansions into the digital collections of the National Library of Wales (Welsh Newspaper Online) and of Australia (Trove). The poster will also demonstrate the means by which reprint branching can be mapped using chronology and character clustering and the relative precision of manual and computer-aided techniques. Finally, it will explore the nature of multi-scalar analysis and how we might best reintegrate 'boutique' periodical research, such as the author's *Scissors and Paste Database*, into large-scale text-mining projects.

Bibliography

Adamic, L. A., Lento, T. M., Adar, E. and Ng, P. C. (2016). Information Evolution in Social Networks,*Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16*, doi: 10.1145/2835776.2835827.

Beals, M. H. (forthcoming). The Role of the Sydney Gazette in Scottish Perceptions of Australia, 1803-1842. In Hinks, J. and Feeley, C. (eds.) *Historical Networks in the Book Trade*. London: Routledge.

Smith, D. A., Cordell, R. and Mullen, A. (2015). Computational methods for uncovering reprinted texts in antebellum newspapers, *American Literary History*, **27**(3): 1–15. doi: 10.1093/alh/ajv029.

Scholarly Research Activities and Digital Tools: When NeMO met FLOSS

Agiatis Benardou
a.benardou@dcu.gr
Digital Curation Unit, Athena Research Centre

Valentine Charles
valentine.charles@europeana.eu
Europeana Foundation

Nephelie Chatzidiakou
n.chatzidiakou@dcu.gr
Digital Curation Unit, Athena Research Centre

Panos Constantopoulos
p.constantopoulos@dcu.gr
Digital Curation Unit, Athena Research Centre; Athens University of Economics and Business

Costis Dallas
c.dallas@dcu.gr
Digital Curation Unit, Athena Research Centre; Panteion University; University of Toronto

Ana Isabel González Sáez
anaisabel.gonzalez@mecd.es
MODUL University Vienna

Sergiu Gordea
Sergiu.Gordea@ait.ac.at
AIT - Austrian Institute of Technology GmbH

Lorna M. Hughes
Lorna.hughes@glasgow.ac.uk
University of Glasgow

Themistoklis Karavellas
tkaravellas@beeldengeluid.nl
Netherlands Institute for Sound and Vision

Gregory Marcus
gmarkus@beeldengeluid.nl
Netherlands Institute for Sound and Vision

Leonidas Papachristopoulos
l.papachristopoulos@dcu.gr
Digital Curation Unit, Athena Research Centre

Vayianos Pertsas
Vpertsas@gmail.com
Digital Curation Unit, Athena Research Centre; Athens University of Economics and Business

While there has been a significant investment in the development of digital tools that can be used in the humanities, information about their use is frequently located in disciplinary silos, with little transfer of knowledge about the features of specific tools that make them valuable for research across the humanities. This poster shows the collaboration between two initiatives, the NeDiMAH Methods Ontology (NeMO) and the EuropeanaTech FLOSS Inventory Task Force. The aim was to carry out research in order to align the FLOSS Inventory against the Activity Types in NeMO, the Ontology of Digital Methods for the Humanities developed by the Digital Curation Unit (DCU), IMIS-ATHENA R.C with the ESF Network for Digital Methods in the Arts and Humanities (NeDiMAH).

The FLOSS Inventory is an effort undertaken by the Netherlands Institute for Sound and Vision and EuropeanaTech to raise awareness of, share access to, and improve the overall status of Open Source software available for cultural heritage developers internationally. The Inventory contains over 200 well-documented, active and relevant OS tools and is actively updated and maintained.

NeMO provides a conceptual framework for representing scholarly practice in the Humanities. This is the main output of NeDiMAH, a Network that ran from 2011- 15 and brought into collaboration 16 countries to document the practice of Digital Humanities across Europe in a series of Methodological Working Groups. Building on earlier expertise in digital taxonomies for the digital humanities, NeDiMAH facilitated a research project carried out by DCU, building upon earlier work on scholarly activity modeling in projects including DARIAH, EHRI and Europeana Cloud. NeMO is a formal ontology which enables the representation and codification of scholarly work by providing a controlled vocabulary of interrelated concepts. NeMO offers a flexible tagging system through a taxonomy of Activity Types, structured in five hierarchies that correspond roughly to scholarly primitives (Unsworth, 2000), and incorporates existing taxonomies and related work such as TadiRAH, Oxford ICT, and DH Commons.

The research teams working on FLOSS and NeMO collaborated to map each tool in the FLOSS inventory against NeMO Activity Types. According to the structure established by NeMO, scholarly research practices are divided into five core Activity Types within a scholarly research lifecycle: acquiring, communicating, conceiving, processing and seeking, which encompass narrower terms accounting for further detail and specialization. This study allows the integration of the information about tools gathered by FLOSS into a uniform conceptual framework for expressing knowledge about scholarly work. By doing so, it also validates the ability of the NeMO ontology to act as a sound framework for the conceptual representation of digital tools and services in one important area of the humanities. This mapping enables the categorisation of available tools according to the function they serve, and could permit researchers in the Humanities - even those without a technical background - to consult an authoritative list of tools covering their needs according to the type of activity they wish to undertake. Availability and the role that each tool can play in the research practice may increase its overall use by the community. The representation of the FLOSS Inventory using NeMO adds value to digital research, and the visualization of categorization ratios it provides facilitates an important debate about software development trends, probing the question whether development is weighted towards software tools that address researchers' needs, a major topic of research in Research Infrastructures across Europe and beyond.

Bibliography

Benardou A., P. Constantopoulos, C. Dallas (2013). An approach to analyzing working practices of research communities in the humanities. *International Journal of Humanities and Arts Computing*, 7: 105-27.

Hughes, L. M. (2011a). *Digital Collections: Use, Value and Impact*. London: Facet.

Hughes, L. M. (2011b). ICT Methods for digital collections research. In Hughes, L. M. (ed.), *Digital Collections: Use, Value and Impact*. London: Facet.

Hughes, L. M. (2014). Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In Schreibman, S. and Siemens, R. (eds), *The New Companion to Digital Humanities*. Oxford: Blackwell.

The AHRC Methods carried out scoping studies on the challenges of access to digital tools in the Humanities: see the series of Working Papers on Digital Tools for the Arts and Humanities: http://www.methodsnetwork.ac.uk/resources/workingpapers.html, and an expert seminar on digital tools: http://www.methodsnetwork.ac.uk/redist/pdf/wg1report.pdf

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this, *Humanities Computing, Formal Methods, Experimental Practice Symposium*, pp. 5-100.

Reflecting On And Refracting User Needs Through Case Studies In The Light Of Europeana Research

Agiatis Benardou
a.benardou@dcu.gr
Digital Curation Unit, ATHENA R.C., Greece

Alastair Dunning
alastair.dunning@theeuropeanlibrary.org
Europeana Foundation

Stefan Ekman
stefan.ekman@snd.gu.se
University of Gothenburg

Vicky Garnett
garnetv@tcd.ie
Trinity College Dublin

Caspar Jordan
caspar.jordan@gu.se
University of Gothenburg

Ilze Lace
ilze.lace@gu.se
University of Gothenburg

Eliza Papaki
e.papaki@dcu.gr
Digital Curation Unit, ATHENA R.C., Greece

Summary: This poster presents work on documenting user needs in the Humanities and Social Sciences as illustrated through Case Studies in the context of the Europeana Cloud “Unlocking Europe's Research via the Cloud” project. Conducted as part of a wider methodological effort including desk research, expert fora and a web survey, methodology and findings of actual use of innovative digital tools and services will be visually represented. This work will form the basis of the Europeana Research Case Studies which will seek to gather and process an evidence-based record of the information practices, needs and scholarly methods in the respective communities.

This poster reports on collaborative, cross-European work conducted during 2013-2015 in the context of the Europeana Cloud “Unlocking Europe's Research via the Cloud” project, and touches upon planned activities in the context of Europeana Research in 2016. Europeana Research is an initiative which aims to create stronger links between the cultural heritage sector and academia. More particularly, it aims to ensure that open, high-quality data from the cultural sector is available for reuse by the digital humanities community.

One of the main objectives of Europeana Cloud was the enhancement of the understanding of digital tools, research processes and scholarly content used in the Humanities and Social Sciences, thus informing the development of tools and aggregation of content in Europeana for research purposes. To this end, and in order to contribute towards the development of the new platform of Europeana Research, Case Studies were developed as part of a wider methodological effort which included desk research, expert fora and web survey for reaching user requirements.

The purpose of this poster is to visually represent the methodology followed and results reached in documenting actual use of innovative digital tools and services in the Humanities and Social Sciences research communities illustrated in three main Case Studies in the disciplines of Education, Art History and Sociology, and further complemented by satellite cases.

The Case Studies were initially selected based on the disciplines and tools that might best make use of current Europeana content. By defining “innovative” as “either performing functions that were previously unavailable, or performing already available functions in a qualitatively different way”, three tools were identified as best fitting this criteria (Transana, HyperImage, NodeXL) enriched by two “satellite” tools more frequently used in the respective research disciplines (NVivo, Voyant).

These were further approached following a threefold methodology of semi-structured interviews, empirical observation of the tools and background research. The results were then discussed both from the perspective of the discipline area, and through the lens of the scholarly primitives (Unsworth 2000, Palmer et al 2009), to determine their use with Europeana content. The poster will also highlight the importance of accessibility of data for research infrastructures and research groups and need to focus on high quality metadata and content both for Europeana Research and the wider GLAM sector, and will illustrate how digital tools are not themselves a guarantee of good research, as researchers do not necessarily use the same digital tool throughout the research process; rather they use one tool per step (one tool = one research primitive).

This poster will also present future work planned to be undertaken in the context of Europeana Research in 2016. Based on Europeana Cloud, a series of new Case Studies will be developed and expanded towards different research communities. The Europeana Research Case Studies will be undertaken in collaboration with existing European research initiatives, and will seek to gather and process an evidence-based record of the information practices, needs and scholarly methods of arts and humanities and social sciences researchers within the broad Europeana ecosystem and particularly in relation to Europeana content. The Case Studies will employ a mixed methods approach combining various ways of gathering empirical evidence on the information needs and scholarly methods employed in digitally-enabled arts and humanities and social sciences research across Europe and beyond.

Bibliography

Hughes, L. (2011). Using ICT methods and tools in arts and humanities research, L. Hughes (Ed.), *Evaluating and measuring the value, use and impact of digital collections*. London: Facet Publishing.

Palmer, C. L., Teffeau, L. C. and Pirmann, C. M. (2009). *Scholarly information practices in the online environment*. Report commissioned by OCLC Research. Published online at: www.oclc.org/programs/publications/reports/2009-02.pdf.

University of Virginia (2005). *Summit on Digital Tools for the Humanities - Report on Summit Accomplishments*. Retrieved from http://www.iath.virginia.edu/dtsummit/SummitText.pdf.

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. In J. Unsworth (Ed.), *Humanities Computing, Formal Methods, Experimental Practice Symposium*, pp. 5-100.

Unsworth, J. (2003). Tool-Time, or Haven't We Been Here Already?. Presented at the *Transforming Disciplines: The Humanities and Computer Science*. Washington, DC. Retrieved from http://people.lis.illinois.edu/~unsworth/carnegie-ninch.03.html.

The symogih.org Project: Towards an International Consortium

Francesco Beretta
francesco.beretta@ish-lyon.cnrs.fr
CNRS - Université de Lyon, France

Vincent Alamercery
vincent.alamercery@ens-lyon.fr
CNRS - Université de Lyon, France

Djamel Ferhod
djamel.ferhod@ish-lyon.cnrs.fr
CNRS - Université de Lyon, France

symogih.org, a project in constant evolution

The collaborative platform for historical research developed as part of the *symogih.org* project has reached a mature stage[1]. With an active SPARQL endpoint, the system is now interoperable and interconnected[2]. It is housing 16 ongoing projects, including several on a European scale (France, Germany, Switzerland, Belgium); there are about 50 active users and it contains nearly 1.700.000 data rows. The research projects include, for example, the Siprojuris project[3], a prosopographic data publishing site interconnected with IdRef, the authority data repository of the French higher education libraries' catalogue[4]; the Historical Atlas of Political Territories[5], a project that is mapping the evolution of world political borders; Society Religion Science (SRS)[6], an experimental digital intellectual history website providing documents encoded in XML following the recommendations of the *Text Encoding Initiative* (TEI) and annotated using the *symogih.org* ontology[7].

The institution hosting the project is the *Laboratoire de recherche historique Rhône-Alpes*[8]. Its human and material resources being limited, to avoid a growth crisis and with a view to building an international community of users, the evolution of two principle aspects of the architecture of the *symogih.org* project needs to be reconsidered and planned for from today: these are,its interface with other information systems, in particular with those put in place by heritage institutions; and the management of a growing community of users.

The purpose of our DH 2016 Conference poster is to share our prospects with other DH projects and to present a proposal for the evolution of the platform and operational information, for both the technological and project management domains, and to call for an international collaboration which could function as a consortium.

Towards a distributed and multi-instantiated database

A possible technological answer to the challenge of the increasing number of projects hosted by the *symogih.org* platform, i.e. to provide the expected information centralization without creating a bottleneck and low performance, is to evolve its architecture by distributing the data between a primary node and several secondary nodes of a clustered database.

The primary node would contain all of the shared repositories (object authority records and instances of the collaborative ontology), as well as general-interest data. The secondary nodes, hosted in various research institutions, would contain the information produced by the local projects. To allow the collaborative management of a shared, good-quality ontology to be maintained, avoiding redundant and messy data, a user of any secondary node shoud be able to access all the data, whether it is held in their database instance or not, via a table held in the primary node, which would hold the keys of all the existing knowledge units and their location in the distributed system.

The whole information system will be interoperable with reference ontologies like CIDOC-CRM or FRBRoo and interlinked with authority files repositories (ISNI, VIAF, BNF, etc.). In addition, a growing portion of the information will be directly accessible in RDF format through a SPARQL endpoint, allowing the data to be queried at the same time as those from other linked data warehouses. The data will be available under a Creative Commons 4.0 international licence.

Establishing an international community of users

This distributed platform management entails the creation of an international organized community of users in the form of a consortium. This community will need to be built around a governance committee, responsible for ensuring a collaborative approach at all levels, from managing the definitions of ontology instances through to data capture. The primary node could be managed, as at present, by the *Pôle histoire numérique* (Digital History Department) of the LARHRA, founding institution of the *symogih.org* project; but it could be hosted in the data centre of an institutional structure such as, in France, the TGIR HumaNum[9].

The other nodes would be managed by teams made up of researchers and IT technicians, ensuring not only management of the respective databases and software development for the whole project, but also running local scientific projects and training users in data modelling and capture, working closely with the governance committee. The evolution towards the participation of projects using different languages will imply the implementation of a multilingual version of the information system, both at the level of the interface and for drafting instances of the collaborative ontology.

By offering access to a collaboratively designed ontology and a robust technological solution, intended for an increasingly wide community of users, the *symogih.org* project will contribute to the evolution of practices in the domain of data production and data curation in historical research.

Bibliography

Beretta, F. and Butez, C. (2014). SyMoGIH project and Geo-Larhra: A method and a collaborative platform for a digital historical atlas. *Digital Humaities 2014 Conference*, Lausanne, Switzerland. http://dharchive.org/paper/DH2014/Paper-831.xml.

Beretta, F., Ferhod, D., Gedzelman, S. and Vernus, P. (2014). The SyMoGIH project: publishing and sharing historical data on the semantic web. *Digital Humanities 2014 Conference*. Lausanne, Switzerland, http://dharchive.org/paper/dh2014/poster-930.xml ; https://halshs.archives-ouvertes.fr/halshs-01097399v1 .

Beretta, F., Vernus, P. and Hours, B. (2012). The Système Modulaire de Gestion de l'Information Historique (SyMoGIH): a collaborative and cumulative platform for storage and use of geo-historical information. *Digital Humanities 2012 Conference*. Hamburg, Germany, http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/le-systeme-modulaire-de-gestion.1.html.

Notes

[1] *SyMoGIH: Système Modulaire de Gestion de l'Information Historique* - Modular Historical Information Management System. The project site: http://www.symogih.org/. Cf. Beretta, F., Vernus, P., and Hours, B. (2012).

[2] http://www.symogih.org/?q=rdf-publication . Cf. Beretta, F., Ferhod, D., Gedzelman, S. and Vernus, P. (2014).

[3] http://siprojuris.symogih.org/

[4] http://www.idref.fr/.

[5] http://geo-larhra.ish-lyon.cnrs.fr/?q=atlas-historique-des-territoires-politiques. Cf. Beretta, F. andButez, C. (2014).

[6] http://srs.symogih.org/.

[7] https://groupes.renater.fr/wiki/symogih/symogih_manuel/edition_de_textes_en_xml-tei .

[8] LARHRA CNRS UMR 5190 – Universités de Lyon et Grenoble:http://larhra.ish-lyon.cnrs.fr/.

[9] Très Grande Infrastructure de Recherche (very large research infrastructure): http://www.huma-num.fr/ .

Crosslingual Textual Emigration Analysis

Andre Blessing
andre.blessing@ims.uni-stuttgart.de
University of Stuttgart, Germany

Jonas Kuhn
jonas.kuhn@ims.uni-stuttgart.de
University of Stuttgart, Germany

The presented work describes the adaptation of a Natural Language Processing (NLP) based biographical data exploration system to a new language. We argue that such a transfer step has many characteristic properties of a typical challenge in the Digital Humanities (DH): Resources and tools of different origin and with different accuracy are combined for their use in a multidisciplinary context. Hence, we view the project context as an interesting test-bed for a few methodological considerations.

In previous work, we developed a web-based application called Textual Emigration Analysis (TEA) (Blessing and Kuhn, 2014). The system consists of two components. The import component automatically extracts emigration paths from a German Wikipedia data set. The user interface component provides several views (interactive map, aggregated tables and underlying textual content) on the extracted data. The whole application was originally designed for German Wikipedia articles and the applied NLP pipeline is based on webservices of the CLARIN-D infrastructure (Mahlow et al., 2014). Later, other German sources of biographical data were included by adapting the import component (ÖBL: Austrian Biographical Dictionary 1815-1950, and NDB: the New German Biography).

One often requested feature was the adaptation of the system to other languages. In DH research it is important to investigate different sources. As a consequence, textual data may include different languages. In particular, biographical analysis systems benefit if sources of different languages can be analysed. However, the development of such language-sensitive systems still lacks sufficient support. Therefore, the used methods should not require any knowledge of the new target language during development phase. In this work we present the adaptation process for French.

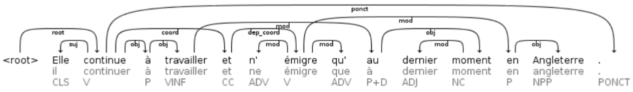
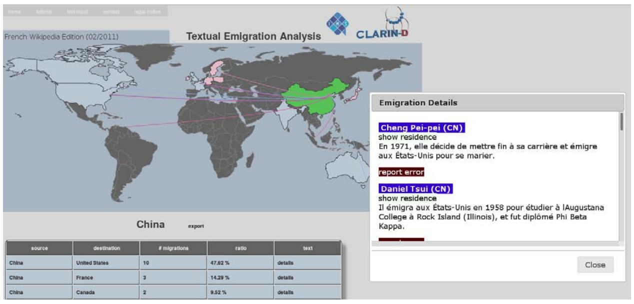


Figure 1: Parsed French sentence which describes an emigration path.Wikipedia and Wikidata are important resources for the development of language technology tools which also holds for cross-lingual approaches. Wikidata enables a clean mapping between the different language editions of Wikipedia. Following the idea of cross-lingual distant supervision (Blessing 2012), our method consists of three steps. First, we use the results of our TEA tool to find biographical articles that include mentions of emigration paths. Subsequently, Wikidata is used to map those articles to their corresponding French articles, if they exist. Finally, we use anchor points in the text to find comparable sentences.

In most cases, emigration sentences contain geospatial and time expressions (e.g., "He emigrated in 1941 to Switzerland."), which can be used to find comparable sentences in the target language. We exploit the hyperlink structure of Wikipedia to recognize geospatial expressions and HeidelTime (Strötgen, 2013) to identify time expressions. The locations can be mapped through Wikidata into the target language and the atomic date representation of HeidelTime enables a simple identification of all matching sentences in the target sentence.


French Wikipedia Edition (02/2011)
Textual Emigration Analysis
CLARIN-D
China
export
Emigration Details
Close

Cheng Pei-pei (CN)
show residence
En 1971, elle décide de mettre fin à sa carrière et émigre aux États-Unis pour se marier.
report error

Daniel Tsui (CN)
show residence
Il émigra aux États-Unis en 1958 pour étudier à lAugustana College à Rock Island (Illinois), et fut diplômé Phi Beta Kappa.

source	destination	# migrations	ratio	text
China	United States	10	47.62 %	details
China	France	3	14.29 %	details
China	Canada	2	9.52 %	details

Figure 2: Screenshot of the TEA web applicationThis results in an annotated corpus in the new target language which can be used as training data for the emigration extraction component . Each sentence of the training corpus is automatically enriched with linguistic annotations (Figure 1) which is necessary to extract features for the emigration extraction component.

Figure 2 depicts our web-based application after integrating the automatically learned French emigration data. Our system can be accessed online: http://clarino1.ims.uni-stuttgart.de/tea

Bibliography

Blessing, A. andSchüthze,H. (2012). Crosslingual Distant Supervision for Extracting Relations of Different Complexity. *Proceedings of the twenty-first ACM International Conference on Information and Knowledge Manageme2012 (CIKM-12),* ACM, New York, NY, USA.

Blessing, A. and Kuhn, J. (2014). Textual Emigration Analysis (TEA). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) European Language Resources Association (ELRA)*, Reykjavik, Iceland.

Mahlow, C., Eckart, K., Stegmann, J., Blessing, A., Thiele, G., Gärtner, M. and Kuhn, J. (2014). Resources, Tools and Applications at the CLARIN Center Stuttgart. *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014).*

Strötgen, J. and Gertz, M. (2013). Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*. Springer **47**(2): 269-98.

Detecting Musical Paratext at Scale

Mark Edward Boettcher
mark.e.boettcher@dartmouth.edu
Dartmouth College, United States of America

John Wallace
john.m.wallace@dartmouth.edu
Dartmouth College, United States of America

Summary

In this work we will detail the open source tools, custom code and processing steps that were used to support textual analysis for a Digital Humanities (DH) project. Additionally, a more general genre classification tool will be described which was developed to support a machine learning model. Extensions of this work will be proposed that may allow these tools to identify types of text.

Background

The Multimedia in the Long 18th Century (Wallace et al., 2015) project is an attempt to automate the process of detecting and quantifying music references in French and English manuscripts written in the 'Long 18th Century', a period ranging from 1685 to 1815. Hundreds of thousands of volumes have been scanned to digital image format and are readily available from a number of online sources. The process of automatically analyzing these images and identifying music rendered in standard music notation is relatively trivial. However, the vast majority of music in the corpus is represented as text, often with various keywords or key phrases, i.e. 'sing the following to the tune of…'. Most people can recognize the difference between poetry, lyrics and prose almost immediately. We assimilate and analyze a large number of clues and usually reach the correct conclusion without conscious effort. This is not the case with the current state of the art in computer technology. While many separate pieces of the solution already exist, gaps still exist. To achieve our goals, it was decided that it would be necessary to leverage the existing work of others while adding a few innovations of our own.

Methods

We have developed a number of batch-processing tools to retrieve the relevant scanned PDF files from online sources and split them into individual images, each representing one page. We use the open source ImageMagick image processing tools to clean up the images. These are then fed to a customized version of the Tesseract OCR engine. Tesseract performs the optical character recognition (OCR) to convert the source images to HOCR files containing bounding box, plain text and recognition confidence levels for each individual paragraph, line and word. This is used to produce PDF files that duplicate the page layout of the original image and contain searchable text that is typically >90% accurate. We have developed an in-house tool to create training data for the machine learning (ML) algorithms used later on. The tool has a user interface that displays both the original image and the formatted PDF side by side. This tool allows the user to choose a category such as: lyrics, poetry, key word, key phrase, etc. The user is able to select regions of the page encompassing anywhere from a single word to the entire page. This generates a very fine-grained database, accurate to within a single word. This database is subsequently used to train the ML engine and to perform validation of the automated classification algorithms. Similar solutions typically have the users classify sections by operating on the page after it has been processed by the OCR engine. This can lead to misclassifications when the OCR results have numerous misspellings or imprecise page layout formatting. We have the user perform the manual classification on the original page images. This provides precise control and future proofing against improvements in OCR techniques. We employ a multi-pronged, multi-pass approach in order to differentiate music lyrics from other types of verse or prose. We perform detailed shape analysis to identify sections of text that 'look' like some form of verse. Symmetry, right and left indentation, relative length of alternating lines and other shape factors are examined. We look at letter casing based on the OCR-processed version. As a result we achieve 90-100% correct identification of structured verse for almost every volume with a relatively small number of false positives. Once we have identified sections of verses, we used a number of empirically determined factors such as relative word frequency, occurrence of n-grams, proximity of keywords/key phrases and comparison to databases of known song lyrics.

Future Development

We are continuing to fine-tune our model, especially in our ability to handle poorly scanned or otherwise nonconforming documents. All software and techniques that we have developed will be made publicly available so that we can share the results of our efforts and with the hope that others will feel free to contribute. Beyond music detection, we see a number of potential applications in the Digital Humanities. The training set region selection can be used just as easily to select features in cuneiform script or Norse runes.

Bibliography

Wallace, J., Sanders, S. and Boettcher, M. (2015). Multi-Media in the Long Eighteenth Century, In *DLFM'15 Proceedings of the 2nd International Workshop on Digital Libraries for Musicology*, pp. 29-32.

Collaborative Annotation and Exploration of Literary Works in Learning Contexts

Andrea Bolioli
andrea.bolioli@celi.it
CELI, Italy

Riccardo Tasso
tasso@celi.it
CELI, Italy

In this poster we present the web application "CBook" (or "Crunched Book"), i.e. a DH tool for collaborative reading, annotation and exploration of literary works. The web app can be found here: https://cbook.it/ . It was used in learning contexts during the research projects *Librare* (in high schools) and *Pinocchio: la comunità dei balocchi* (in middle schools). Other teachers are using it in complete autonomy in their classrooms.

Students and teachers can annotate portions of text, explore novel's settings through interactive maps, add and share images and other contents related to a literary work, search for terms into the text (e.g. in the quoted speeches of a particular character), explore connections between characters in their social networks.

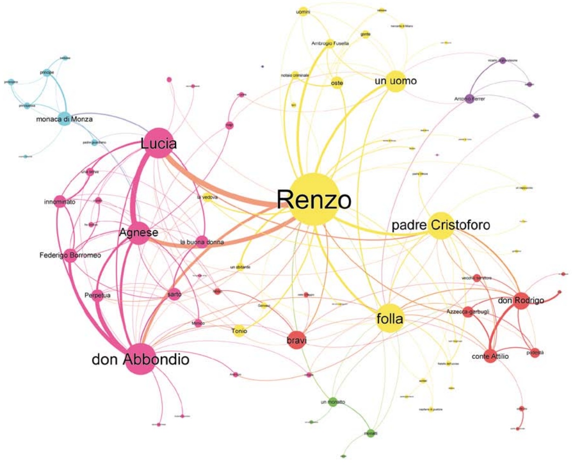
The CBook is used by italian students and their teachers, as an innovative digital tool for reading and studying literature, in primary and secondary schools. The list of schools involved in *Librare* project, for example, can be found here:

Some Italian and English classics are already present in the web app, the full work or a portion: *I promessi sposi* (complete), *Le avventure di Pinocchio* (complete), *Decameron* (3 novels), *Odissea* (book VI, italian translation), *Romeo and Juliet* (complete), *The Wonderful Wizard of Oz* (chapters 15). Other works requested by teachers (e.g. Dante's *Purgatorio*), will be added to the web app in the next months.

We can say that CBook is, in some sense, a collaborative anthology (or an Anthology 2.0), i.e. a collection of literary works selected by the users that are reading, annotating and exploring the works, using some DH tools and methodologies (collaborative annotation, interactive maps, SNA of characters, text mining, etc).

The web app was designed and tested in collaboration with high school teachers, students and DH scholars. It is used on tablets, PCs, interactive whiteboards, and smartphones (with less functionalities).

An important feature of CBook is interoperability between models and tools for TEI XML annotation and models and tools for semantic annotation. A significant difference between the CBook and other annotation tools is its emphasis on user-centered design (for students and teachers).



The annotation system of CBook is based on Annotator (http://annotatorjs.org/) , "an opensource JavaScript library to easily add annotation functionality to any webpage", following annotation standards for digital documents developed by the W3C Web Annotation Working Group. The User Interface is written in HTML5, CSS and Javascript. The texts were annotated in XML, using a simplified version of TEI. On the server side, we use a graph Database and Java.

The actions of cbook's users (sign in, sign out, comment) can be monitored in a public dashboard accessible here: http:/sensori.librare.org/librare/. In this dashboard you can see the events concerning both cbooks (digital) and paper books that users were working on in this project.

The first version of "The digital lab of crunched books" was the Second Runner Up of Digital Humanities Awards 2014 in the Best DH Tool or Suite of Tools section (3 march 2015): http://dhawards.org/dhawards2014/results/

Bibliography

Bolioli, A. and Tasso R. (2014). Enjoying classic literature, in *Interactive e-Books for Children IBooC2014* (Denmark), 2nd Workshop at IDC Interaction Design and Children.

Bolioli, A., Casu M., Lana M. and Roda R. (2013). Exploring the Betrothed Lovers, in M. Finlayson, B. Fisseni, B. Lowe and J. C. Meister (eds.), *2013 Workshop on Computational Models of Narrative*, OASIC, vol. **32**.

Pierazzo, E. (2015). *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate Publishing, Ltd.

Verbal Identity of a Fictional Character: a Quantitative Study with a Machine Learning Experiment

Anastasia Bonch-Osmolovskaya
abonch@gmail.com
National research university 'Higher school of economics', Moscow, Russia

Elena Sidorova
sieleny@gmail.com
National research university 'Higher school of economics', Moscow, Russia

Daniil Skorinkin
skorinkin.danil@gmail.com
National research university 'Higher school of economics', Moscow, Russia

Introduction

The idea that narrative literature comprises two distinct types of speech – that of the narrator and that of characters – dates as far back as Plato's *Republic*. The philosopher distinguished **diegesis** (narrative, narration), when the author speaks for himself, from **mimesis** (imitation, enactment), when the author puts words in the mouths of his or her fictional actors.

Modern narratology places high emphasis on the concepts of **point of view** (POV) and **POV structure** of a text (Schmid, 2003), which are often expressed through specific combinations of author's and character's speech. Authors may switch between the POV's by employing character-specific lexica and the use of temporal and spatial references that indicate certain POV (Uspensky, 1983).

Leo Tolstoy was one of the writers known for **conscious** usage of such means to differentiate between the character POV's. He was a firm proponent of the idea that each character has to speak his/her **own** language if the book was to be convincing. Critics confirm that Tolstoy's characters do have their personal styles of conversation.

In this paper we made an attempt to provide quantitative grounds for these claims. For that purpose we extracted all speech activity instances from *War and Peace*, attributed them to the speaker characters and used the data to train a classifier. Our hypothesis was that if Tolstoy's characters actually possessed these unique speech features, the classifier would be able to predict the speaker with some tolerable accuracy.

Data

Instances of direct speech were extracted from the text with help of ABBYY Compreno (Starostin et al., 2014). For more details on the extraction procedure see (Bonch-Osmolovskaya A., Skorinkin D., 2015). The total number of extracted speech instances was 6853, of which 4476 had their speakers identified.

Apart from the speaker, a number of additional attributes were extracted for each instance: text of the speech, text of the author's introduction ('she cried', 'he said with a laugh'), normalized speech predicate ('to say', 'to cry', 'to whisper', 'to burst out'), the number of question and exclamatory phrases within one speech, the number of words in the speech and the number of punctuation marks.

Before we carried out the experiment we attempted to analyze the data and detect potentially informative features. It appeared that certain characters (Natasha Rostov, Nikolai Rostov) tend to speak in short intermittent bursts and exclaim a lot, so they were expected to have higher average punctuation marks per word ratio and bigger shares of exclamatory speech. To confirm this intuition we gathered some aggregated statistics (see Fig. 1).

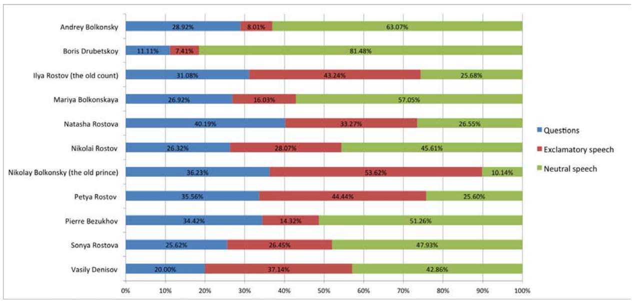


Fig. 1 Shares of exclamatory and question sentences in the speech of the main characters

Exclamatory and question phrases together make up the 'emotional part' of a character's speech. Its share (Fig. 2) seems to correlate with age extremities. Prince Nikolay Bolkonsky is probably the oldest of the main characters, and as his age gets the better of him in the course of the novel, he turns more and more emotional and impulsive; Petya Rostov, on the other hand, is an exuberant and emotional boy, the youngest of the Rostov family.

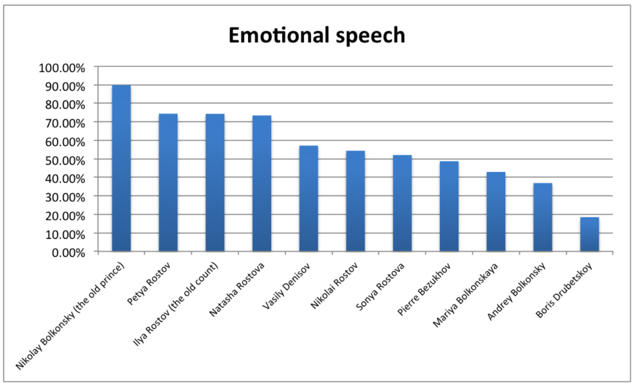


Fig. 2 Characters with the highest share of 'emotional speech' (exclamatory and question sentences combined)

Fig. 3 reflects character's overall punctuation marks per word ratios. Seems like the 'burst speech' pattern is hereditary within the Rostov family:

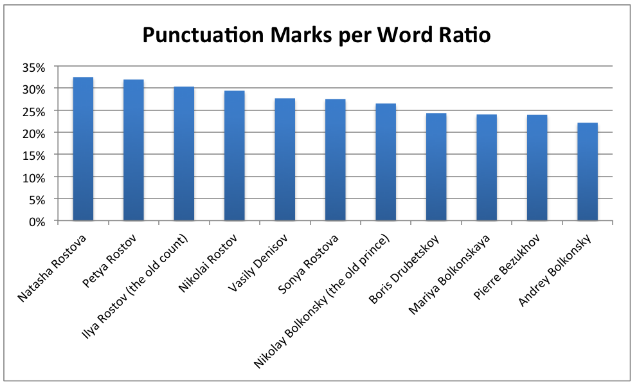
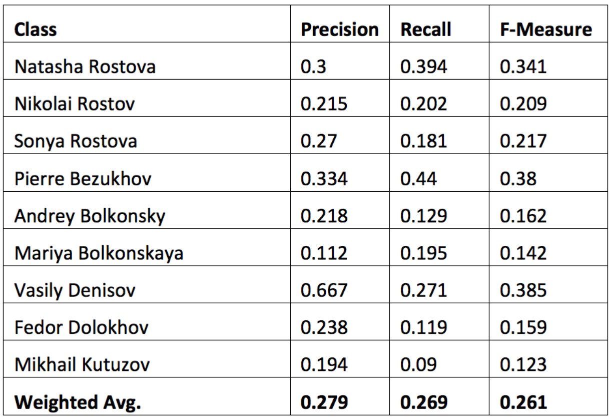


Fig. 3 Punctuation marks per word ratio in the direct speech text

Machine learning experiment

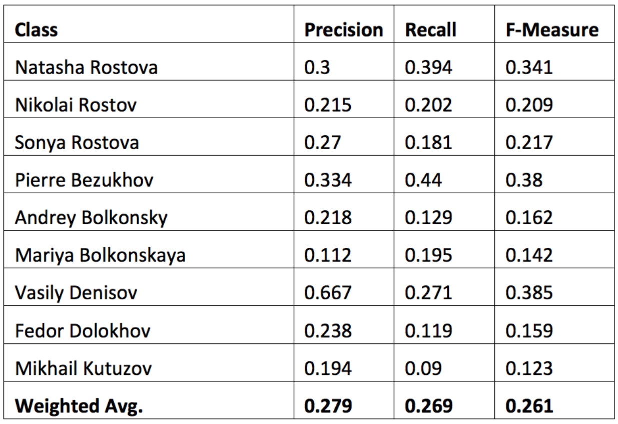
Next step was to try and use some of these features to train a classifier. We used several standard algorithms, of which Random Forest demonstrated the best results. At the first stage we created a baseline by training the classifier solely on the lemma and word form frequencies of speech. Table 1 shows the results we obtained.



Class	Precision	Recall	F-Measure
Natasha Rostova	0.3	0.394	0.341
Nikolai Rostov	0.215	0.202	0.209
Sonya Rostova	0.27	0.181	0.217
Pierre Bezukhov	0.334	0.44	0.38
Andrey Bolkonsky	0.218	0.129	0.162
Mariya Bolkonskaya	0.112	0.195	0.142
Vasily Denisov	0.667	0.271	0.385
Fedor Dolokhov	0.238	0.119	0.159
Mikhail Kutuzov	0.194	0.09	0.123
Weighted Avg.	**0.279**	**0.269**	**0.261**

Table 1 Baseline results for classifier trained on lemma and word form frequencies

The second stage was to add formal features that we considered informative (number of exclamatory phrases, number of questions and punctuation marks per word ratio) and retrain the classifier. The results we obtained show that the use of these features slightly improved performance.



Class	Precision	Recall	F-Measure
Natasha Rostova	0.3	0.394	0.341
Nikolai Rostov	0.215	0.202	0.209
Sonya Rostova	0.27	0.181	0.217
Pierre Bezukhov	0.334	0.44	0.38
Andrey Bolkonsky	0.218	0.129	0.162
Mariya Bolkonskaya	0.112	0.195	0.142
Vasily Denisov	0.667	0.271	0.385
Fedor Dolokhov	0.238	0.119	0.159
Mikhail Kutuzov	0.194	0.09	0.123
Weighted Avg.	**0.279**	**0.269**	**0.261**

Table 2 Results for classifier trained with additional features

Results & discussion

Our first attempts to automatically classify speaker in Tolstoy's text did not prove successful. The best F-measure we were able to obtain so far does not exceed 0.385 for an individual character. However, we were able to show that some formal features, such as punctuation marks per word ratio or the number of exclamatory/question sentences, might improve classification quality. This assumption can be confirmed by the figures in the Data section, where the aggregated values of features correspond with certain character traits that are apparent to the human reader.

Bibliography

Bonch-Osmolovskaya A., Skorinkin D. (2015). Automatic semantic tagging of Leo Tolstoy's works. In Abstracts of Digital Humanities – 2015 conference, Sydney, Australia. http://dh2015.org/abstracts/xml/SKORINKIN_Daniil_Automatic_semantic_tagging_of_Le/SKORINKIN_Daniil_Automatic_semantic_tagging_of_Leo_Tols.html (accessed 06 March 2016)

Eichenbaum, B. (2009). Works on Leo Tolstoy. Saint-Petersburg. SPBSU Faculty of Philology and Arts.

Schmid, V. (2003). Narratology. Moscow: LRC Publishing House.

Starostin A. , Smurov I., Stepanova M. (2014). A Production System for Information Extraction Based on Complete Syntactic-Semantic Analysis. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference 'Dialogue', Bekasovo, pp. 659–667.

Uspensky, B. (1983). A Poetics of Composition: The Structure of the Artistic Text and Typology of a Compositional Form. Oakland: University of California Press.

Mobile Makerspaces: Te(a)chnology, Design and Digital Humanities

Christina Boyles
christina-boyles@uiowa.edu
University of Iowa, United States of America

Lindsay Kistler Mattock
lindsay-mattock@uiowa.edu
University of Iowa, United States of America

This project explores the development of a mobile makerspace for graduate and undergraduate DH scholars at the University of Iowa and Grinnell College. Blending the approaches of makerspaces like the DHMakerBus <http://dhmakerbus.com> and University of Victoria's Maker Lab <http://maker.uvic.ca/>, this DH makerspace will investigate the use of a suite of tools designed to support the development digital literacy and technological proficiency for students across the DH curriculum. By combining the mobility of the DHMakerBus with the experimental computing of the University of Victoria's Maker Lab, we will produce a new method of digital humanities pedagogy that welcomes the participation of primary and secondary students and educators, local citizens, and digital humanities practitioners.

Designed to support experiential learning - learning through making - in undergraduate and graduate classrooms, across campus, and across Iowa, the makerspace focuses on the computer processes that go into making rather than the products produced by them. As such, it emphasizes the process of making itself, whether successful, failed, or flawed.

In *Debates in the Digital Humanities* Alexander Reid suggest that most graduate students have had little exposure to digital technology during their undergraduate education, "enter[ing] his or her graduate education as a novice in regards to the digital" (357). It is likely, therefore, that his assessment applies to our broader community of Eastern Iowa. To address the our users' lack technological expertise, we are scaffolding projects for learners—starting with littlebits, moving to RaspberryPi, and ending with Aruduino. These tools were selected for the pilot study for their mobility, accessibility, and affordability. Each of these factors makes it easy to implement and replicate the makerspace in unconventional venues—secondary schools, adult education classes, and community events. Additionally, these tools gradually increase participants' comfort and literacy with digital tools.

Te(a)chnology

This mobile DH makerspace utilizes three "gateway" technologies selected for their robust design, sophisticated abilities, and ease of use, and portability:

1. *littleBits* - a digital building kit based on the logic of Lego that allow students and makers to build digital tools and explore the internal logic of computers <http://littlebits.cc>

2. *Raspberry Pi* - a small, but powerful computer, about the size of a credit card with a GUI interface and Python encoding that integrates touchscreens and digital cameras <https://www.raspberrypi.org>

3. *Arduino Kit for littleBits* - build on the capabilities of the littleBits library by adding the Arduino computer, teaching students how to expand their coding skills by working with Java-based programming <http://littlebits.cc/kits/arduino-coding-kit>

Together these three technologies scaffold to develop users' confidence in coding, building, project management, and digital literacy. This particular suite of tools is easily transportable to classrooms and campuses, or may be shipped to students participating in online sections of workshops and seminars. Furthermore, the mobility of the makerspace allows for collaborations with non-traditional and non-academic communities including secondary students, adult education seminars, community partners, and interested citizens. Participants will come away from workshops utilizing this mobile makerspace with increased knowledge of computer processes, an awareness of design thinking, and the confidence to collaborate on a variety of projects with diverse teams.

Methodology

This poster will report on the results of a pilot study of the mobile DH makerspace. Participants in the pilot study included undergraduate and graduate students and faculty completing design challenges and product tests. Working with a convenience sample of graduate and undergraduate students, the makerspace was tested in two modes: (1) students working together in the same space and (2) students working across a distance, collaborating virtually through videoconferencing and other digital collaboration tools.

We assessed the increased digital literacy and confidence of participants through a series of surveys, visualization exercises, focus group interviews, and participant observation. The results of this pilot study will inform the development of individual course units and workshops for courses within the DH Curriculum at the University of Iowa and Grinnell College, in addition to workshops for faculty and staff across both campuses and outreach efforts to reach the off-campus community.

Conclusions

This mobile makerspace affords opportunities for makers to build confidence creating and experimenting with digital tools while allowing students to build digital literacy skills and address other key aspects of DH education, including: working in interdisciplinary teams, applying digital practices, managing projects, and explaining technology (Rockwell and Sinclair 182-183). Ultimately, we argue that this model can be emulated in other educational settings as a new model for DH pedagogy that is more accessible and more collaborative than traditional makerspaces.

A special thanks to Miriam Posner for generously agreeing to review this abstract and provide feedback and suggestions.

Bibliography

Reid, A. (2012). Graduate Education and the Ethics of the Digital Humanities. In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 350-67.

Rockwell, G. and Sinclair, S. (2012). Acculturation and the Digital Humanities Community. In Hirsch, B. D. (ed), *Digital Humanities Pedagogy: Practices, Principles, and Politics*. Cambridge: Open Book Publishers, pp. 177-211.

Studying Linguistic Changes on 200 Years of Newspapers

Vincent Buntinx
vincent.buntinx@epfl.ch
EPFL (École polytechnique fédérale de Lausanne),
Switzerland

Cyril Bornet
cyril.bornet@epfl.ch
EPFL (École polytechnique fédérale de Lausanne),
Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
EPFL (École polytechnique fédérale de Lausanne),
Switzerland

Newspaper archives as a linguistic corpus

This research investigates methods to study linguistic evolution using a corpus of scanned newspapers. We use a corpus of 4 million press articles covering about 200 years of archives, thus documenting indirectly the evolution of written language. The corpus is made out of digitized facsimiles of *Le Journal de Genève* (1826–1997) and *La Gazette de Lausanne* (1804–1997). For each journal, the daily scanned issues were algorithmically transcribed using an OCR system. The whole archive represents more than 20 TB of scanned data and contains about two billion words, putting it beyond the capabilities of most usual analysis techniques for regular desktop computers.

The corpus can be easily divided into subsets corresponding to the year of publication. However, the number of pages and their content fluctuates greatly depending on the year, ranging from 280'000 in the early 19th century to about 18 million in the later years of the 20th century. Figure 1 shows the relative size of each subset in terms of number of words for *Le Journal de Genève* (JDG) and *La Gazette de Lausanne* (GDL).

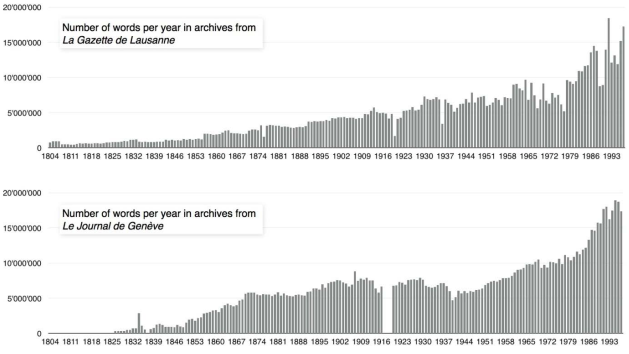


Figure 1: corpus size versus years for GDL (top) and JDG (bottom).

Considering the lack of data for *Le Journal de Genève* for the years 1837, 1917, 1918 and 1919, we left those out in all further graphs and analytics. In addition, some years had to be removed because the scanning quality was too poor (1834, 1835, 1859 and 1860 for JDG and 1808 for GDL).

Lexical kernels: Definition and basic measures

A straightforward approach to the problem consists in computing a textual distance between subsets of the corpora. One could, for instance, easily compute the so-called Jaccard distance (Jaccard 1901, Jaccard 1912) between two consecutive vocabularies. In the same way, other distances could also be tried, such as those given by Kullback and Leibler (1951), Kullback (1987), Chi-squared distance (Sakoda, 1981), and Cosine similarity (Singhal, 2001).

However, the uneven distribution of the corpus subsets (Figure 1) causes methodological difficulties for interpreting these distances. An increase in the lexicon size causes an indirect increase in the linguistic drift as measured by the Jaccard formula (Sternitzke and Bergmann, 2009). Under such conditions, it is difficult to untangle the effects of the unevenness of the distribution of subsets of the corpus from the actual appearance and disappearance of words.

These difficulties of interpretation motivate the exploration of another, possibly sounder approach to the same problem. Let us define a *lexical kernel* $K_{x,\, y,\, C}$ as the sequential subset of unique words common to a given period starting in year *x* and finishing in year *y* of a corpus *C*.

$K_{1804,1998›GDL}$ is, for instance, the subset of all words present in the yearly corpus of *La Gazette de Lausanne*. It contains 5242 unique words that have been used for about 200 years. The kernel $K_{1826,1998,JDG}$ contains 7486 unique words, covering a period of about 170 years. As the covered period is smaller, the kernel is naturally larger.

The exact contents of both $K_{1804,1998,GDL}$ and $K_{1826,1998,JDG}$ are provided in the appendix. It is interesting to note that 4465 words are in common between the two kernels.

Figure 2 shows the statistical distribution of word-typologies for both kernels.

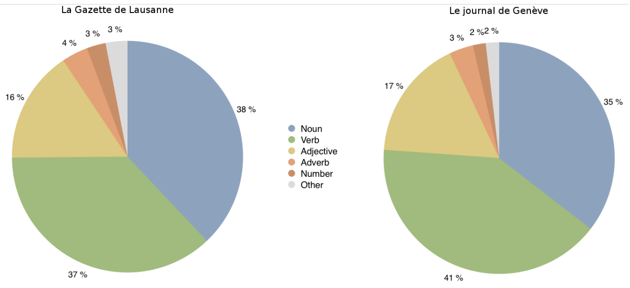


Figure 2: Distribution in terms of typologies of words contained in the kernel of $K_{1798,1998,\mathrm{GDL}}$ (left) and $K_{1826,1998,\mathrm{JDG}}$ (right).

Word resilience

Extending the notion of a kernel, it is rather easy to study the resilience of a given word. Let R_d be the union of all words contained in a kernel corresponding to a duration of $y - x \geq d$ years. For instance, $R100$ contains all the words that maintain themselves in the corpus for at least 100 years. R subsets are organized as concentric sets:

$$R_1 \subset R_2 \subset \dots \subset R_i \subset R_i + 1$$

The relative proportion of each subset sheds light on both the stability and dynamics of language change. Figure 3 shows the distribution of word resilience for both journals.

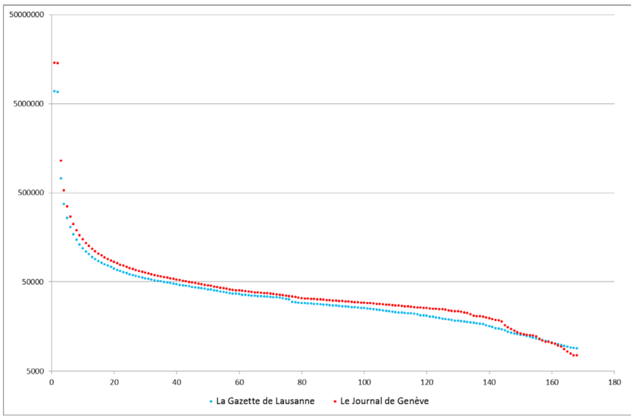


Figure 3: Size of R_d versus the number of maintained years d (logarithmic scale) showing the word resilience distribution for JDG (green) and GDL (blue).

The GDL resilience curve is normalized (on the same years range as JDG) in order to make the two curves comparable. This representation of R_d shows a similar global word resilience trend for both JDG and GDL. However, we notice that the two curves intersect when considering the longest durations.

Discussion

Large databases of scanned newspapers open new avenues for studying linguistic evolution. However, these studies should be conducted with sound methodologies in order to avoid misinterpretation of artifacts. Common pitfalls include misinterpreting results linked to the size variation of the subsets or overgeneralizing results obtained on one particular newspaper corpus to general linguistic evolution.

In this paper, we have introduced the notion of a kernel as a possible approach to study linguistic changes under the lens of linguistic stability. Focusing on stable words and their relative distribution is likely to make interpretations more robust.

Results were computed on two independent corpora. It is striking to see that most of the results obtained are extremely similar for both. The kernels composition in terms of grammatical word typologies is very similar. Results in terms of word resilience are also similar. This suggests that our methods are indeed measuring general linguistic phenomena beyond the specificity of the corpora chosen for this study. However, this still needs to be confirmed with subsequent studies involving other corpora, such as non-journalistic texts and texts in other languages.

Bibliography

Jaccard, P. (1901). Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles 01/1901*, **37**(142): 547–79.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, **11**: 37–50.

Kullback, S. (1987). Letter to the Editor: The Kullback-Leibler distance. *The American Statistician*, **41**(4): 340–41.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**(1): 79–86.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**(8): 707–10.

Sakoda, J. M. (1981). A Generalized Index of Dissimilarity. *Demography*, **18**(2): 245–50.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, **24**(4): 35–43.

Sternitzke, C. and Bergmann, I. (2009). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, **78**(1): 113–30.

Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie Analysis

Manuel Burghardt
manuel.burghardt@ur.de
Media Informatics Group, University of Regensburg, Germany

Michael Kao
Michael.Kao@stud.uni-regensburg.de
Media Informatics Group, University of Regensburg, Germany

Christian Wolff
christian.wolff@ur.de
Media Informatics Group, University of Regensburg, Germany

Introduction: Quantitative movie analysis

Film studies make use of both, qualitative as well as quantitative methods (Korte, 2004). While there is a large variety of qualitative approaches to analyze movies (cf. e.g. Monaco, 2009; Sikov, 2010), most quantitative attempts seem to be focused on the analysis of the length and frequency of a film's shots, which are understood to be the "single definable elements which can be nominated and described" (Salt, 2006: 14). After Barry Salt's[1] seminal article "Statistical Style Analysis of Motion Pictures" appeared in 1974, numerous other quantitative studies were to follow[2]. "Cinemetrics" (Tsivian, 2009) has been suggested as a term to describe these quantitative, shot-based approaches for analyzing movies. Cinemetrics is also the name of a large online database that contains information about shot lengths and frequencies for several thousand films[3].

Studies that take into account quantifiable parameters other than shots are, however, rather rare. Among the few examples are Hoyt et al. (2014), who describe a tool that can be used to visualize relations between the characters of a film. Another example can be found in Ewerth et al. (2009), who present a toolkit that allows researchers to automatically detect shots and camera motion, superimposed text, faces and audio signals. While the latter example is rather focused on the automatic annotation of quantitative features, other projects, such as Lev Manovich's (2013) „Visualizing Vertov", focus on the presentation and visualization of quantitative parameters.

In this paper, we suggest to enhance the existing, shot-focused approaches to quantitative movie analysis, by considering additional parameters, such as language (cf. Forchini, 2012) and color use (cf. Flückinger, 2011). We present a prototype that can be used to automatically extract and analyze these parameters from movies and that makes the results accessible in an interactive visualization.

A prototype for the analysis of language and color use in movies

Language use as well as the use of colors in movies have long been known in the area of qualitative film studies. We argue, that these parameters are equally suited for a quantitative approach and present an experimental prototype that can be used to quantify the language and color used in different movies. Much like Clement et al. (2008), who discuss the drawbacks and opportunities of computer-based methods in the field of literary studies, we believe that digital, quantitative movie analysis tools can be helpful in "offering provocations, surfacing evidence, suggesting patterns and structures, or adumbrating trends". As the prototype allows researchers to investigate potential correlations between color usage and corresponding language in a movie, it can be used to examine questions such as the following:

- Are there characteristic patterns in color or language use for movies from different eras, genres, or directors (e.g. dark colors and words such as "kill" or "blood" in horror movies)?
- Are there characteristic color patterns within a film that correlate with the occurrence of certain characters or objects (e.g. bright colors whenever the hero is speaking)?
- Are there characteristic color patterns within a film that correlate with the sentiment of the language (e.g. dark colors for language with negative sentiment)?

Obtaining language and color data

Machine-readable instances of movie language can be obtained fairly easy in the form of subtitle files, which are freely available via sites such as OpenSubtitles[4]; for a precompiled corpus of subtitles also cf. Tiedemann (2012). The standard file structure of such subtitles contains a timestamp as well as a transcription of the actual dialog fragments.

Information about color usage can be extracted directly from the movie itself, by cutting the digital movie into single frames and by calculating the most frequent colors for each frame (color histograms).

Analyzer component

Our prototype comprises an analyzer and a viewer component. The analyzer can be used to extract single frames from a movie by using the open source tool FFmpeg[5]. We used a K-means Cluster algorithm (Wu, 2012) to group together similar RGB values in each frame, as the actual variation of distinct RGB values is too high to allow for any kind of meaningful, quantitative interpretation. The analyzer also processes the subtitle file of a movie and uses Python NLTK[6] to perform basic POS tagging, as we are mainly interested in how nouns correlate with certain colors. We used the Python library TextBlob[7] to perform a simple sentiment analysis for each of the adjectives, tagging them with a polarity score between -1 (negative) and +1 (positive). After the analysis, each frame is saved as a JPG file; all quantitative data is stored in a JSON file.

Viewer component

The viewer component uses this data to generate an interactive HTML page that can be viewed in any recent web browser. A popular visualization of the most frequent colors that occur in a movie can be found in the MovieBarcodes[8] project. In a MovieBarcode, each frame of a movie is skewed to be only one pixel wide; all frames are then lined up in a row that looks very much like a colored barcode. On the overview page of our tool, all movies that have been analyzed before are rendered in a MovieBarcode visualization, together with information of the four most frequent colors in the movie (cf. Fig. 1). This view can be used to compare various movies with each other from a more distant perspective.

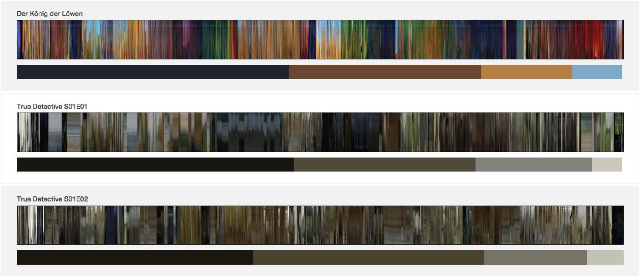


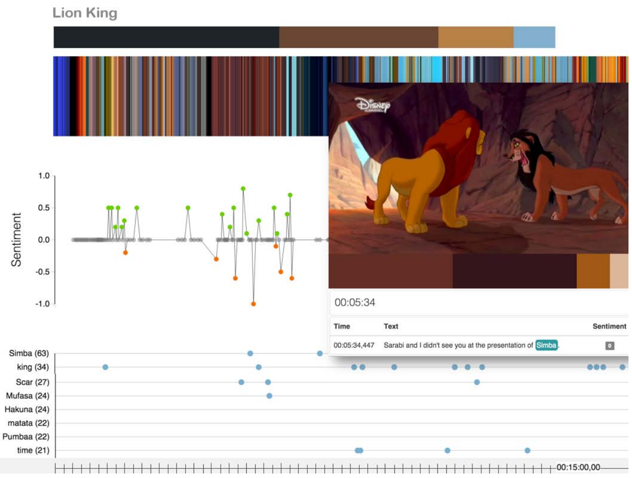
Figure 1: Overview of analyzed films in a MovieBarcode visualization ("The Lion King", top; "True Detective, season 1, episode 1", middle; "True Detective, season 1, episode 2", bottom).

By clicking on one of the MovieBarcodes, the tool zooms into the respective movie and renders different kinds of information in a more detailed view (cf. Fig. 2).

Below the MovieBarcode appears a sentiment graph that aggregates a score between -1 and +1 for each dialog. In the bottom row, the most frequent nouns are displayed. All different types of information are also aligned to the time axis of the movie. The visualization is fully interactive, i.e. by hovering over one of the frames in the MovieBarcode, or a node in the sentiment graph or the noun distribution, the corresponding frame and subtitle appear as an overlay. The complete movie can also be navigated back and forth by means of the arrow keys.

Conclusion and future directions

We believe the real strength of a quantitative approach that makes use of language and color information lies in a mix of "distant watching" (cf. Howanitz, 2015) and close watching, i.e. characteristic language-color patterns identified in specific movies can be used as a query to search other movies for similar patterns. Our next steps will therefore go into the direction of an information system that allows researchers to search and compare a collection of movies according to language and color characteristics.



Time	Text	Sentiment
00:05:34,447	Sarabi and I didn't see you at the presentation of Simba.	0

Figure 2: Detailed view with MovieBarcode, sentiment graph and noun distribution.

Bibliography

Buckland, W. (2008). What Does the Statistical Style Analysis of Film Involve? A Review of Moving into Pictures. More on Film History, Style, and Analysis. *Literary and Linguistic Computing*, **23**(2): 219-30.

Flückiger, B. (2011). Die Vermessung ästhetischer Erscheinungen. *Zeitschrift für Medienwissenschaft*, **5**(2): 44–60.

Forchini, P. (2012). *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Bern et al.: Peter Lang Verlag.

Clement, T., Steger, S., Unsworth, J. and Uszkalo, K. (2008). How not to read a million books. http://people.brandeis.edu/~unsworth/hownot2read.html (accessed 3 March 2016).

Hoyt, E., Ponot, K. and Roy, C. (2014). Visualizing and Analyzing the Hollywood Screenplay with ScripThreads. *Digital Humanities Quarterly*, **8**(4). http://www.digitalhumanities.org/dhq/vol/8/4/000190/000190.html
(accessed 3 March 2016).

Howanitz, G. (2015). Distant Waching: Ein quantitativer Zugang zu YouTube-Videos. *Digital Humanities im deutschsprachigen Raum (Dhd) 2015: Conference Abstracts*. Graz, pp. 33-39. http://gams.uni-graz.at/o:dhd2015.abstracts-gesamt (accessed 3 March 2016).

Korte, H. (2004). *Einführung in die Systematische Filmanalyse*. Berlin: Erich Schmidt Verlag.

Manovich, L. (2013). Visualizing Vertov. *Softwarestudies.com*. http://softwarestudies.com/cultural_analytics/Manovich.Visualizing_Vertov.2013.pdf (accessed 3 March 2016).

Monaco, J. (2009). *Howto Read a Film: Movies, Media, and Beyond*. Oxford (NY): Oxford University Press.

Salt, B. (2006). *Movinginto Pictures. More on Film History, Style, and Analysis*. London: Starword Publishing.

Salt, B. (1974). Statistical Style Analysis of Motion Pictures. *Film Quarterly*, **28**(1): 13-22.

Sikov, E. (2010). *Film Studies. An Introduction*. New York: Columbia University Press.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC) 2012*. Istanbul, pp. 2214-18.

Tsivian, Y. (2009). Cinemetrics, Part of the Humanities' Cyberinfrastructure. In Ross, M., Grauer, M. and Freisleben, B. (eds.), *Digital Tools in Media Studies – Analysis and Research. An Overview*. Bielefeld: tanscript Verlag, pp. 93-100.

Ewerth, R., Mühling, M., Stadelmann, T., Gllavata, J., Grauer, M. and Freisleben, B. (2009). Videana: A Software Toolkit for Scientific Film Studies. In Ross, M., Grauer, M. and Freisleben, B. (eds.), *Digital Tools in Media Studies – Analysis and Research. An Overview*. Bielefeld: tanscript Verlag, pp. 100-16.

Wu, J. (2012). *Advances in K-means Clustering: A Data Mining Thinking*. Berlin and Heidelberg: Springer-Verlag.

Notes

[1] For a concise review of Barry Salt's work on quantitative movie analysis cf. Buckland (2008).

[2] For a comprehensive overview of Cinemetrics-related research cf. the bibliography compiled by Mike Baxter, available online at http://www.cinemetrics.lv/dev/bibliography_with_essay_Baxter.pdf (accessed 3 March 2016).

[3] http://www.cinemetrics.lv (accessed 3 March 2016).

[4] www.opensubtitles.org (accessed 3 March 2016).

[5] https://www.ffmpeg.org/ (accessed 3 March 2016).

[6] http://www.nltk.org/ (accessed 3 March 2016).

[7] http://textblob.readthedocs.org/ (accessed 3 March 2016).

[8] http://moviebarcode.tumblr.com (accessed 3 March 2016).

From Ashes to Ashé: Digital Disaster Archives as Memorials

Patricia Lynn Carlton
carltonpatricia@knights.ucf.edu
University of Central Florida, United States of America

The rapid proliferations of digital archives that develop in response to, and often in the immediate aftermath of, globalized, catastrophic events are not only repositories for historical documentation and collective memories, but are also spaces where rituals of mourning and formations of subject identity can be critically examined. This poster presents close and distanced readings of the mission statements and selected verbal and visual elements of four disaster archives/memorials: The September 11 Documentary Project; The National September 11 Memorial Museum; Our Marathon: The Boston Bombing Digital Archive and WBUR Oral History Project; and UC QuakeStudies.

Optimally, the disaster memorial archive aggregates and displays artifacts of mourning, recovery, and a diversity of shared memories while simultaneously facilitating a sense of public agency and participation in the community's collective wellbeing. As evidenced by the expansion of local cultural heritage archives and crowdsourced collection development, collections of local, private memories co-exist with other documents of official or journalistic origins. Archivists Joan Schwartz and Terry Cook are among many who discuss the paradigm shifts in the archive profession that have been initiated by postmodern ideas and facilitated by the affordances of digitization (Schwartz and Cook, 2002). Contemporary archives reach into the local constituencies and beyond in the increasingly global and interdisciplinary network of digital archives (McKemmish and Gilliland, 2004: 84).

Notwithstanding the stated missions of disaster archives to provide an open-source archive that facilitates the participation by all constituents and interested public, the archival affinity for historical truth and broad representation of its public can sometimes work against public mourning and recovery. As one of my research questions, I explore whether the archival affinity for historical truth and broad representation of its public can co-exist with public mourning and recovery. Modern archives were traditionally associated with progressive and positivist concepts of history whereas memory was associated with unstable sensory triggers and artifices (including rhetorical applications). Postmodern influences and recordings of 20th and 21st century witnesses to and survivors of war crimes and other atrocities have influenced the policies and heuristics used by archivists and historians to critically assess and incorporate individual and collective memories of previously marginalized populations in the archive.

Although skepticism towards the evidentiary value of collective memory has abated, the tasks of balancing memories with official accounts are, nevertheless, complicated. The eventual selection, classification, and creation of policies for access and long-term preservation are negotiable products—an outcome of the balance of power between the public and the archive, and also between the digital media and technical platforms. From my research that triangulates data from my critical analysis of the purpose and functionality of disaster archives, and field studies, I coded themes and abstracted dimensions of private and public mourning, formations of subject identity (such as witness, victim, first responder, or "other"), and evidence of ethics and personal judgment used to create archival policies and collections. The refinements of these dimensions and data visualizations are represented in the poster.

Beyond the discovery of relationships between these dimensions and attaining insight into, if not a solution to, whether the historical and memorializing functions can seamlessly merge in the disaster archive, this poster also rationalizes the interdisciplinary methods of ethnographic and social science research as applied to humanities issues and, conversely, assumes a humanities perspective in critiquing the technical infrastructures of the reviewed disaster archive/memorials. The digital disaster archive promotes the active outreach and mobilization of local communities and establishes models for building resilience to the globalized man-made and naturally caused disasters.

The aesthetics of the archive—the database structure and narratives told within—is also an ethical decision. Sharon Daniel (2007: 150) describes these collaborative spaces (archives, in general) as "dialogic spaces in which the acts of writing, imaging, storytelling, and political statement are a collective production, a process rooted in social interaction and dialogue that produces a narrative without authorial consistency." The narratives emerging from the disaster archive/memorials convey an urgency to rebuild and perhaps, redefine the rituals of mourning and civic responsibility to others. It is the intent of this poster to illustrate the narrative arc of my research, highlighting the areas of interest for digital humanists, including the construction of an ethical and aesthetic database, and to illuminate patterns of online memorialization in disaster archives.

Bibliography

Daniel, S. (2007). Database: An Aesthetics of Dignity. In Vesna, V. (ed.) *Database Aesthetics: Art in the Age of Information Overflow*. University of Minnesota Press, pp. 150.

Library of Congress. (2002). September 11, 2001, Documentary Project. [Online] Available at: https://www.loc.gov/collections/september-11th-2001-documentary-project/about-this-collection/ (Accessed 03 March 16).

MacNeil, H. (2012). What finding aids do: archival description as rhetorical genre in traditional and web-based environments. *Archival Science: International Journal on Recorded Information*, Academic OneFile, EBSCOhost, **4**: 485.

Marathon.neu.edu. (2016). Our Marathon. [Online] Available at: http://marathon.neu.edu/ (Accessed 3 Mar. 2016).

National September 11 Memorial and Museum | World Trade Center Memorial. (2015). 9/11 Memorial.org. [Online] Available at: http://www.911memorial.org/. (Accessed 03 March 16).

Roy Rosenzweig Center for History and New Media and American Social History Project/Center for Media and Learning. (2002). *The September 11 Digital Archive: Saving the Histories of September 11, 2001.* [Online] Available at: http://911digitalarchive.org/ (Accessed 03 March 16).

Schwartz, J, and Cook, T. (2002). Archives, records, and power: the making of modern memory. *Archival Science*, Library, Information Science and Technology Abstracts, EBSCOhost, **2**(1-2): 1-19.

UC QuakeStudies | UC CEISMIC Canterbury Earthquakes Digital Archive. (2011). UC QuakeStudies. [Online] Available at: https://quakestudies.canterbury.ac.nz/ (Accessed 03 March 16).

Poetry in Prose: automatic identification of verses in brazilian literature

Ricardo Carvalho
ricardo.sys@gmail.com
State University of Feira de Santana, UEFS

Angelo Loula
angelocl@gmail.com
State University of Feira de Santana, UEFS

João Queiroz
queirozj@gmail.com
Federal University of Juiz de Fora, UFJF

In 1946, the brazilian poet Guilherme de Almeida published a study on the structured patterns of verses that he discovered in the prose of 'Os Sertões' ('Rebellion in the Backlands') by Euclides da Cunha (1902). According to Almeida's work, there is, in the Euclidean prose, apparently more often at the end of the paragraphs, versification structures of various rhythmic patterns. In 1996, another study on the same literary work was published by Augusto de Campos validated Almeida's discovery and revealed several others versified patterns in Euclidean prose. Dodecasyllables and Alexandrines are among the most used metric patterns, in varied combinations and positions. The diversity of patterns found, disregarding "the strict metrification" and admitting "more rhythmic freedom" (Campos, 1996), creates surprising zones of tension, "areas spread with poetry in significant portions of poetry in his prose" (Campos, 1996).

The process of separating and nesting of poetry syllables, the scansion process, is usually applied to text structures categorically defined as poetry, allowing mapping of poetry metric characteristics present in the author's writing. Performed by a person, the same analysis carried out by Almeida and Campos require, depending on the size of the piece, hours, days or even months of work.

Our work proposes the use of computational techniques to perform the process of automated scansion and analysis of Euclides da Cunha's prose, revealing its verse structures, thus reducing time for the task, providing a new tool for prose analysis and opening a new research agenda. These verse structures, distributed along the text, are found using computational methods based on scansion rules for Portuguese language. As the location of these structures are not previously given, any sentence is treated as a potential candidate for a verse, moreover segments of the sentences can also be considered.

In order to identify metric verses in the text, our system performs four major steps: extraction of sentences, separation of syllables, scansion, and overlay of verses in the original text. From a digital copy of the book, sentences are extracted according to punctuation mark present in literary piece. In Portuguese, the rhythm or musicality of a verse follows the alternance of strong (tonic) syllables and weak (atonic) syllables, so along with syllable boundaries, the position of the tonic syllable is also identified for every word. Therefore, every word in each sentence undergoes syllable separation following grammatical rules, applying the software developed by Neto et al (2015), defining initial syllable boundaries. Besides the positions of tonic syllables are also identified, determining rhythmic features.

To identify verses, the final process of scansion is performed, considering intravocabular (syneresis and diaeresis) and intervocabular (elision and crasis) phonological changes. These changes may alter initial syllable count, for example with the omission of one or more sounds, merging two syllables in a single one. As a final output, the verses identified are overlaid on the original document, along with verse classification, metric count, syllable separations and tonic syllables position, replacing the original sentence, allowing analysis by the user in context.

Initial experiments with the proposed system were performed for the book ‘Os Sertões’ by Euclides da Cunha, aiming to reproduce in a computer lab the work performed by Guilherme de Almeida in the 40s and Augusto de Campos in the 90s. As an example of the results, in page 67, the system identified previously twenty-four candidates verses. Of these, we have, "O sertanejo é, antes de tudo, um forte", one segment of text that starts the third chapter, identified by our system as a dodecasyllable "O / ser/ta/ne+/jo é+,/ an+/tes/ de/ tu+/do/ um/ for+/te", where ‘+’ identifies a tonic syllable, and, starting the third paragraph, "É o homem permanentemente fatigado.", a autonomous paragraph which was also identified by our system as a dodecasyllable, “É+ / o ho+/mem / per/ma/nen/te/men+/te / fa/ti/ga+/do.”. Both verses were indicated by Augusto de Campos in his work. Overall, among dodecasyllables and decasyllables, considering only whole sentences as a possible verse, our system was able to identify 273 verses, many of them already manually identified by Almeida or Campos. Nevertheless, previous works on the annotation of verses in the book ‘Os Sertões’ were not comprehensive annotations, so it is not possible to have full statistics on accuracy. Nevertheless, our system has exactly identified 52% of Augusto de Campos’s annotated verses, and the other 48% of the verses were identified with a difference of one unit in syllable count, due to differences in elision.

Bibliography

Almeida, G. (1946). A poesia d’Os Sertões. *Diário de São Paulo*. August 18.

Neto, N., Rocha, W. and Sousa, G. (2015). An open-source rule-based syllabification tool for Brazilian Portuguese. *Journal of the Brazilian Computer Society*, **21**(1): 1-10.

Campos, A. (1996). TRANSERTÕES. *Folha de São Paulo*. November 3.

Towards an XML Corpora Exposition as LOD with the Lightweight Xquery-Based Framework SynopsX

Emmanuel Château-Dutier
emmanuel.chateau.dutier@umontreal.ca
Université de Montréal, Canada

Maud Ingarao
maud.ingarao@ens-lyon.fr
IHPC – UMR 5037

Jean-Philippe Magué
jean-philippe.mague@ens-lyon.fr
Interaction, Corpus, Apprentissages, Représentations – UMR 5191

Philippe Pons
philippe.pons@cnrs.fr
ANR Ampère

Severine Gedzelman
severine.gedzelman@ens-lyon.fr
Laboratoire de recherche historique Rhone-Alpes – UMR 5190; Triangle – UMR 5206

Sylvain Boschetto
sylvain.boschetto@ish-lyon.cnrs.fr
Laboratoire de recherche historique Rhone-Alpes – UMR 5190

Samantha Saidi
samantha.saidi@ens-lyon.fr
Triangle – UMR 5206

Valérie Beaugiraud
valerie.beaugiraud@ens-lyon.fr
Institut d’histoire de la pensée classique – UMR 5037

Carole Boulai
carole.boulai@ish-lyon.cnrs.fr
Triangle – UMR 5206

Pierre-Yves Jallud
pierre-yves.jallud@huma-num.fr
TGIR Huma-Num

Emmanuelle Morlock
emmanuelle.morlock@mom.fr
Histoire et Sources des Mondes Antiques – UMR 5189

Various approaches have been proposed to publish XML/TEI scholarly editions or XML/EAD finding aids on line, but no standard XML software with a ready to use web application has yet emerged. As a possible next step after the XML mark-up of an edition or of an inventory, online publications are still a difficult issue for many projects in Digital humanities. Among available frameworks (like eXist or baseX) there is no ready to use web application to easily expose XML scholarly corpora as Linked Open Data (LOD).

Initiated by the Digital humanities network of ENS Lyon (Atelier des Humanités Numériques de l'ENS de Lyon), SynopsX is a lightweight framework which goal is to easily publish and expose XML scholarly corpora. It's a full XQuery web application built with the native XML database BaseX. Involving different partners from various institutions the project is developed as a free and open source software under GNU, and is hosted on GitHub (https://github.com/synopsx). Three principles have guided the conception of the software: collaboration, mutualization and genericization. Thus, we decided to put together all the publication problems and needs we could encounter in our very various Digital Humanities corpora (disciplines, approaches, institutions) to specify the needs of a generic web application.

SynopsX has been conceived as a scalable and easily customizable solution for XML publication of XML files (TEI, EAD, OAI, etc.). The software brings a templating system for various renderings of XML resources according to predefined or customized mappings from XML data to various output formats. SynopsX use the BaseX implementation of Adam Retter's proposed specifications for RESTXQ. Because it allows full control on the URL scheme to build real REST applications, it could be used to expose XML corpora as Linked Open Data.

This poster will present the needs of LOD for XML corpora publication. It will explain the different strategies that could be used and how we're planning to use a RESTXQ controller with views and models to propose a tool that allows the contents exposition as LOD through a restful API.

Enhancing Close Reading

Muhammad Faisal Cheema
faisal@informatik.uni-leipzig.de
Leipzig University, Germany

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

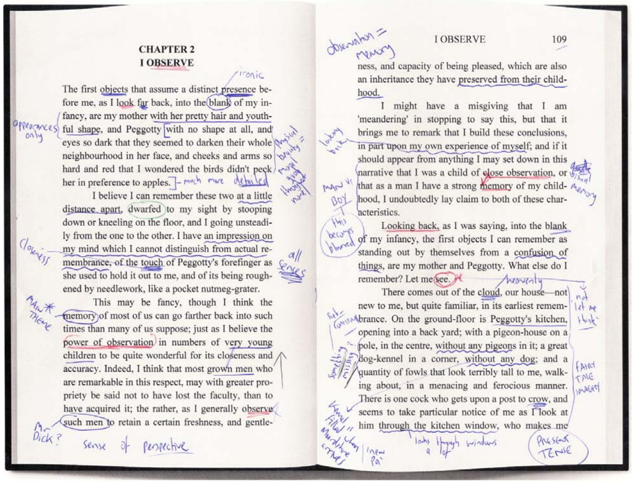
Gerik Scheuermann
scheuermann@informatik.uni-leipzig.de
Leipzig University, Germany

Motivation

In last years, the advancements in computer science brought a global change in the way information is stored, retrieved and analyzed. The digital humanities also benefit from these developments, and now, a vast amount of texts is available in digital form. This information explosion generates interesting research questions for humanities scholars who are capable of deriving new insights from this knowledge bank. In order to support humanities scholars, many visualization techniques – summarized in a survey (Jänicke et al., 2015b) – were developed to aid exploring large texts collections. Most of these techniques are interactive and belong to the category of distant reading (Moretti, 2005). The authors of the mentioned survey observe that less work has been done to improve the close reading capabilities of humanities scholars even though they are often focused on close reading text passages.

Close reading is the careful interpretation of the text, where the scholar iteratively reads the text in order to explore its meaning, inherent topics and occurring relationships (Boyles, 2013). Traditionally, close reading is done on paper. Several ideas and thoughts are made persistent by annotations written at the margins alongside the text (see Figure 1). But as the margin space is limited, not all observations can be put around the text. So, annotations may become cluttered and confusing for the reader, especially, when obsolete ideas are struck through. Despite its disadvantages, annotating on paper is still quite popular as it benefits the scholars to record observations about the hypothesis and all these changes reappear in front of the scholar's eyes as soon as he re-reads the text passage. We observed that the way of annotating in close reading resembles the idea of mind maps (Buzan et al., 1993) that are based on a central concept and thoughts are represented around it using lines and text. In the close reading scenario, the text can be considered as the central concept and annotations represent thoughts.

An important task of computer science is to enhance the original workflows of researchers with computational methods. As most humanities scholars are well trained in close reading and nowadays often work with digital texts, it is necessary to enhance their capabilities for digital close reading. We propose an enhanced close reading design inspired by mind-maps that not only mimics the traditional way of annotating a text on paper, but also helps humanities scholars to perform live visual analyses. Furthermore, we use extendible margins to provide enough space for all thoughts of the scholar.



CHAPTER 2
I OBSERVE

The first objects that assume a distinct presence before me, as I look far back, into the blank of my infancy, are my mother with her pretty hair and youthful shape, and Peggotty with no shape at all, and eyes so dark that they seemed to darken their whole neighbourhood in her face, and cheeks and arms so hard and red that I wondered the birds didn't peck her in preference to apples.

I believe I can remember these two at a little distance apart, dwarfed to my sight by stooping down or kneeling on the floor, and I going unsteadily from the one to the other. I have an impression on my mind which I cannot distinguish from actual remembrance, of the touch of Peggotty's forefinger as she used to hold it out to me, and of its being roughened by needlework, like a pocket nutmeg-grater.

This may be fancy, though I think the memory of most of us can go farther back into such times than many of us suppose; just as I believe the power of observation in numbers of very young children to be quite wonderful for its closeness and accuracy. Indeed, I think that most grown men who are remarkable in this respect, may with greater propriety be said not to have lost the faculty, than to have acquired it; the rather, as I generally observe such men to retain a certain freshness, and gentle-

ironic

appearances only

physical beauty = moral goodness throughout novel

much more detailed

closeness

all senses

MAIN* THEME

Mr Dick?

sense of perspective

I OBSERVE 109

ness, and capacity of being pleased, which are also an inheritance they have preserved from their childhood.

I might have a misgiving that I am 'meandering' in stopping to say this, but that it brings me to remark that I build these conclusions, in part upon my own experience of myself; and if it should appear from anything I may set down in this narrative that I was a child of close observation, or that as a man I have a strong memory of my childhood, I undoubtedly lay claim to both of these characteristics.

Looking back, as I was saying, into the blank of my infancy, the first objects I can remember as standing out by themselves from a confusion of things, are my mother and Peggotty. What else do I remember? Let me see.

There comes out of the cloud, our house—not new to me, but quite familiar, in its earliest remembrance. On the ground-floor is Peggotty's kitchen, opening into a back yard; with a pigeon-house on a pole, in the centre, without any pigeons in it; a great dog-kennel in a corner, without any dog; and a quantity of fowls that look terribly tall to me, walking about, in a menacing and ferocious manner. There is one cock who gets upon a post to crow, and seems to take particular notice of me as I look at him through the kitchen window, who makes me

observation = memory

looking back

MAN vs BOY

this becomes blurred

visual memory

heavenly

not let me think

sub-conscious

something missing?

FAIRY TALE IMAGERY

Kennel filled = when Murdstone comes

new Pa

looks through windows a lot

PRESENT TENSE

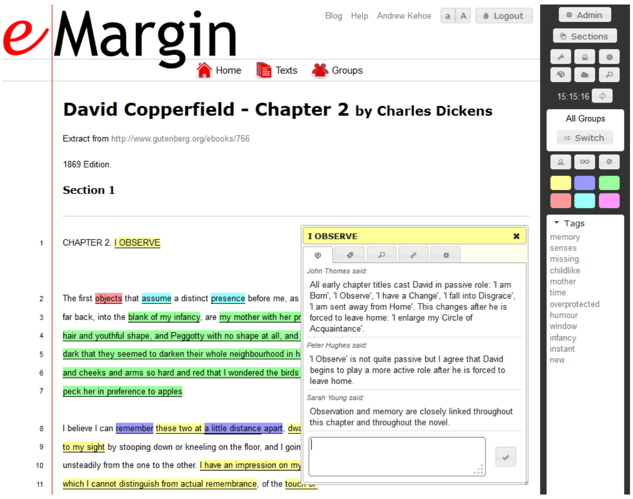
Figure 1: Traditional close reading on paper[1]

Related Work

Nancy Boyles (Boyles, 2013) defines close reading, which has become a fundamental method in literary criticism in the 20th century (Hawthorn, 2000), as follows: "Essentially, close reading means reading to uncover layers of meaning that lead to deep comprehension." Annotating the text in close reading is a strong method for scholars to facilitate the understanding of a text passage. Figure 1 shows the result of a traditional close reading approach. In this example, various annotation methods were used by the scholar to annotate various features of a text passage in Charles Dickens' „David Copperfield".

The availability of digital texts has further awaken the interest of humanities scholars in collaboratively close reading the same texts. There are several annotation tools for such a purpose, such as eMargin (Kehoe et al., 2013), Hypothes.is (Bonn et al., 2014) and NB (Zyto et al., 2012). These tools are beneficial for collaborative research and classroom environments as they provide an excellent paradigm to share thoughts, as well as find collective answers. To avoid clutter, these tools work with popup windows that are only shown on demand. In Figure 2, the eMargin system is shown where colors are used to highlight different text features, and a popup window on demand, lists the comments of collaborating scholars.

Digital Ink Annotations systems (Schilit, 1998, Bargeron et al., 2003, Agrawala et al., 2005, Yoon et al., 2013) also support annotating text, but their use is only limited to pen-based computing devices such as tablets. The systems are designed to work well on smaller screens, and the adaption to larger screens is not appropriately implemented.



David Copperfield - Chapter 2 by Charles Dickens

Extract from http://www.gutenberg.org/ebooks/766

1869 Edition.

Section 1

1 CHAPTER 2. I OBSERVE

2 The first objects that assume a distinct presence before me, as
3 far back, into the blank of my infancy, are my mother with her pr
4 hair and youthful shape, and Peggotty with no shape at all, and
5 dark that they seemed to darken their whole neighbourhood in h
6 and cheeks and arms so hard and red that I wondered the birds
7 peck her in preference to apples.

8 I believe I can remember these two at a little distance apart, dwa
9 to my sight by stooping down or kneeling on the floor, and I goin
10 unsteadily from the one to the other. I have an impression on my
11 which I cannot distinguish from actual remembrance, of the touch of

I OBSERVE

John Thomas said:

All early chapter titles cast David in passive role: 'I am Born', 'I Observe', 'I have a Change', 'I fall into Disgrace', 'I am sent away from Home'. This changes after he is forced to leave home: 'I enlarge my Circle of Acquaintance'.

Peter Hughes said:

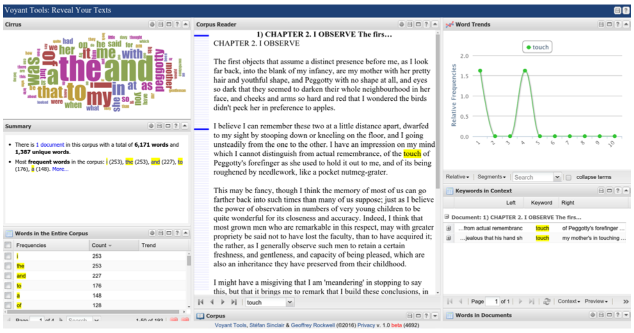
'I Observe' is not quite passive but I agree that David begins to play a more active role after he is forced to leave home.

Sarah Young said:

Observation and memory are closely linked throughout this chapter and throughout the novel.

Figure 2: eMargin annotation tool[2]

Close reading tasks can also be assisted via distant reading tools. For example, parallel coordinates, a heatmap and a dot plot are used to analyze the variance of a selected text passage from different German translations of Shakespeare's Othello (Geng et al., 2013). Heat maps are appropriate visualizations to illustrate the distribution of specific phrases or annotations in a corpus (Muralidharan, 2011, Alex et al., 2015). Voyant Tools allow the user to perform basic text mining functions with selected word statistics shown in linked views (Sinclair et al., 2012). The Voyant Tools interface in Figure 3 shows statistics about Chapter 2 of Oscar Wilde's "David Copperfield". Goffin's idea to enhance close reading is the integration of small visualizations (e.g., maps or bar charts) besides the words of a text (Goffin et al., 2014).



- There is 1 document in this corpus with a total of 6,171 words and 1,387 unique words.
- Most frequent words in the corpus: i (253), the (253), and (227), to (176), a (148). More...

Frequencies	Count	Trend
i	253	
the	253	
and	227	
to	176	
a	148	
of	128	

1) CHAPTER 2. I OBSERVE The firs...

CHAPTER 2. I OBSERVE

The first objects that assume a distinct presence before me, as I look far back, into the blank of my infancy, are my mother with her pretty hair and youthful shape, and Peggotty with no shape at all, and eyes so dark that they seemed to darken their whole neighbourhood in her face, and cheeks and arms so hard and red that I wondered the birds didn't peck her in preference to apples.

I believe I can remember these two at a little distance apart, dwarfed to my sight by stooping down or kneeling on the floor, and I going unsteadily from the one to the other. I have an impression on my mind which I cannot distinguish from actual remembrance, of the touch of Peggotty's forefinger as she used to hold it out to me, and of its being roughened by needlework, like a pocket nutmeg-grater.

This may be fancy, though I think the memory of most of us can go farther back into such times than many of us suppose; just as I believe the power of observation in numbers of very young children to be quite wonderful for its closeness and accuracy. Indeed, I think that most grown men who are remarkable in this respect, may with greater propriety be said not to have lost the faculty, than to have acquired it; the rather, as I generally observe such men to retain a certain freshness, and gentleness, and capacity of being pleased, which are also an inheritance they have preserved from their childhood.

I might have a misgiving that I am 'meandering' in stopping to say this, but that it brings me to remark that I build these conclusions, in

Document: 1) CHAPTER 2. I OBSERVE The firs...

Left	Keyword	Right
...from actual remembranc	touch	of Peggotty's forefinger ...
...jealous that his hand sh	touch	my mother's in touching ...

Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (©2016) Privacy v. 1.0 beta (4692)

Figure 3: Screenshot of web-based Voyant Tools (Sinclair et al., 2012).

Enhanced Close Reading Design

In contrast to the tools mentioned above, we combine traditional annotation tasks with distant reading analyses to enhance the close reading capabilities of the scholar. We suggest a design inspired by mind mapping (an example mind map is shown in Figure 4a), a methodology that allows a researcher to work on a central concept, and thoughts and features about that concept are placed around it using figures, lines etc. In a mind map, the associations spread out from a central concept in a free-flowing, yet organized and coherent manner (Budd, 2004) - thus forming a mental map of the central concept. We observe that like in the case of mind maps, fixed annotations around the central text in a traditional close reading process facilitate forming a mental map of the thoughts about the text of interest, and help the scholar to draw conclusions when seeing the whole picture.

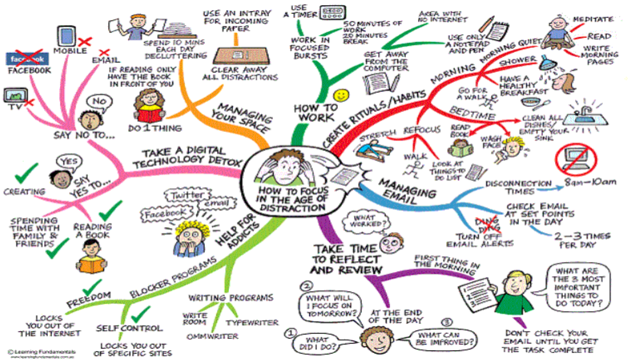

HOW TO FOCUS IN THE AGE OF DISTRACTION
TAKE A DIGITAL TECHNOLOGY DETOX
SAY NO TO...
FACEBOOK
MOBILE
EMAIL
TV
No
SAY YES TO...
YES
CREATING
SPENDING TIME WITH FAMILY & FRIENDS
READING A BOOK
MANAGING YOUR SPACE
SPEND 10 MINS EACH DAY DECLUTTERING
USE AN INTRAY FOR INCOMING PAPER
CLEAR AWAY ALL DISTRACTIONS
IF READING ONLY HAVE THE BOOK IN FRONT OF YOU
DO 1 THING
HOW TO WORK
USE A TIMER
WORK IN FOCUSED BURSTS
50 MINUTES OF WORK 20 MINUTES BREAK
GET AWAY FROM THE COMPUTER
AREA WITH NO INTERNET
USE ONLY A NOTEPAD AND PEN
CREATE RITUALS/HABITS
MORNING
MORNING QUIET
MEDITATE
READ
WRITE MORNING PAGES
SHOWER
HAVE A HEALTHY BREAKFAST
GO FOR A WALK
BEDTIME
CLEAN ALL DISHES/ EMPTY YOUR SINK
READ BOOK
WASH FACE
STRETCH
REFOCUS
WALK
LOOK AT THINGS TO DO LIST
MANAGING EMAIL
DISCONNECTION TIMES
8am–10am
CHECK EMAIL AT SET POINTS IN THE DAY
2–3 TIMES PER DAY
DING DING
TURN OFF EMAIL ALERTS
TAKE TIME TO REFLECT AND REVIEW
WHAT WORKED?
FIRST THING IN THE MORNING
WHAT ARE THE 3 MOST IMPORTANT THINGS TO DO TODAY?
DON'T CHECK YOUR EMAIL UNTIL YOU GET THE TASK COMPLETE
AT THE END OF THE DAY
① WHAT DID I DO?
② WHAT WILL I FOCUS ON TOMORROW?
③ WHAT CAN BE IMPROVED?
HELP FOR ADDICTS
Twitter
email
Facebook
BLOCKER PROGRAMS
FREEDOM
LOCKS YOU OUT OF THE INTERNET
SELF CONTROL
LOCKS YOU OUT OF SPECIFIC SITES
WRITING PROGRAMS
WRITE ROOM
TYPEWRITER
OMMWRITER
© Learning Fundamentals

Figure 4a: An example mind map[3]

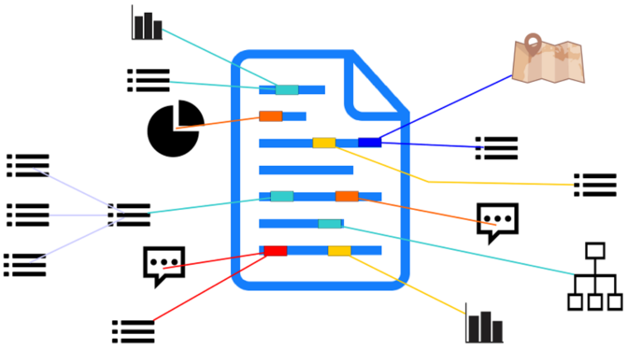
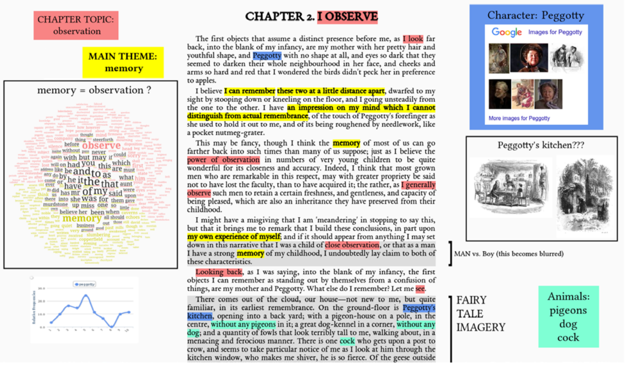


Figure 4b: Mind-map inspired close reading

Figure 4b illustrates the idea of a mind map inspired interface with multiple types of annotations supporting the scholar in the close reading process. Textual annotations known from the traditional close reading are also necessary in the digital process. In addition, images, videos and charts can facilitate text interpretation and the generation of valuable hypotheses about the text. To support dynamic, multifarious views on a certain text passage or a term of interest, we designed our interface the way that the literary scholar can apply a multitude of visual analyses and generate distant reading visualizations that are placed as annotations alongside the text. This combines the traditional close reading paradigm with elaborated text visualization techniques valuable for exploration purposes. An important feature of our proposed interface design is to support the scholar to „stay in the flow" (Bederson, 2004), so that the central focus remains on the text, which can be analyzed without interrupting the scholar. The major advantage of our design over existing tools that assist close reading tasks is interface versatility. For example, Voyant Tools (see Figure 3) provide a predefined set of visualizations based on text statistics. On the other hand, our design allows the scholar to choose an appropriate text visualization as an annotation alongside the text, which is based on a user-defined query on the text.. Therefore, the scholar can apply different text visualizations for different passages of the text to support a variety of close reading tasks.

An example of the design discussed above is shown in Figure 5. The example from Figure 1 is annotated using different kinds of annotations. Like in other digital tools, certain topics of the text are annotated using colors. In addition, the character(s) Peggotty is marked and a panel shows thumbnail images based on a Google Images search. Also the relative word frequency chart of the term "Peggotty" in Chapter 2 is shown on the bottom left. Furthermore, on the left area, a TagPie (Jänicke et al., 2015a) showing the co-occurrences of both the terms memory and observation helps to investigate the hypothesis of the literary scholar about the similar meaning of both topics. The example depicts how the scholar can use different annotation tools as well as different distant reading tools to enrich the close reading experience.



CHAPTER TOPIC: observation

MAIN THEME: memory

memory = observation ?

CHAPTER 2. I OBSERVE

The first objects that assume a distinct presence before me, as I look far back, into the blank of my infancy, are my mother with her pretty hair and youthful shape, and Peggotty with no shape at all, and eyes so dark that they seemed to darken their whole neighbourhood in her face, and cheeks and arms so hard and red that I wondered the birds didn't peck her in preference to apples.

I believe **I can remember these two at a little distance apart**, dwarfed to my sight by stooping down or kneeling on the floor, and I going unsteadily from the one to the other. I have **an impression on my mind which I cannot distinguish from actual remembrance**, of the touch of Peggotty's forefinger as she used to hold it out to me, and of its being roughened by needlework, like a pocket nutmeg-grater.

This may be fancy, though I think the **memory** of most of us can go farther back into such times than many of us suppose; just as I believe the power of observation in numbers of very young children to be quite wonderful for its closeness and accuracy. Indeed, I think that most grown men who are remarkable in this respect, may with greater propriety be said not to have lost the faculty, than to have acquired it; the rather, as I generally observe such men to retain a certain freshness, and gentleness, and capacity of being pleased, which are also an inheritance they have preserved from their childhood.

I might have a misgiving that I am 'meandering' in stopping to say this, but that it brings me to remark that I build these conclusions, in part upon **my own experience of myself**; and if it should appear from anything I may set down in this narrative that I was a child of close observation, or that as a man I have a strong **memory** of my childhood, I undoubtedly lay claim to both of these characteristics.

Looking back, as I was saying, into the blank of my infancy, the first objects I can remember as standing out by themselves from a confusion of things, are my mother and Peggotty. What else do I remember? Let me see.

There comes out of the cloud, our house—not new to me, but quite familiar, in its earliest remembrance. On the ground-floor is Peggotty's kitchen, opening into a back yard; with a pigeon-house on a pole, in the centre, without any pigeons in it; a great dog-kennel in a corner, without any dog; and a quantity of fowls that look terribly tall to me, walking about, in a menacing and ferocious manner. There is one cock who gets upon a post to crow, and seems to take particular notice of me as I look at him through the kitchen window, who makes me shiver, he is so fierce. Of the geese outside

Character: Peggotty

Peggotty's kitchen???

MAN vs. Boy (this becomes blurred)

FAIRY TALE IMAGERY

Animals: pigeons dog cock

Figure 5: Example of our design

Future Work and Conclusions

We held discussion with the collaborating humanities scholars about the design as well as the usability of the proposed interface. The scholars remarked that such an interface will help removing fears of using digital humanities tools and that they intend to use the tool as it mimics their existing workflows. They also mentioned that such a tool could help users getting a better big picture of the text, and that it enhances the close reading capabilities of the scholar. Another important point is the capability in supporting teaching activities. They mentioned that various types of annotations (text, pictures, charts) are also used in teaching material, but it is not easy to share these with students. Such a tool could support this process as it generates persistent annotations to be analyzed and discussed collaboratively in courses.

We observe that the scholar's initial reactions after seeing the prototype of the tool, which is still in development, are convincing and encouraging. We think that rigid modeling syntax is inappropriate for annotation. Our final interface will allow the scholar to make annotation styles versatile. At the digital humanities conference, we will demonstrate our prototype and discuss future prospects within the community. An additional user study will compare the viability of our proposed, mind map inspired annotation technique to existing approaches.

Acknowledgements

We thank our colleagues from the humanities department, Judith Blumenstein in particular, who provided insights and expertise that greatly assisted this research.

Bibliography

Agrawala, M. and Shilman, M. (2005). DIZI: a digital ink zooming interface for document annotation. *Human-Computer Interaction-INTERACT 2005*, Springer Berlin Heidelberg, pp. 69-79.

Alex, B., Grover, C., Zhou, K., Hinrichs and Palimpsest, U. (2015). Improving Assisted Curation of Loco-specific Literature. *Proceedings of the Digital Humanities 2015*, pp. 5-7.

Bargeron, D. and Moscovich, T. (2003). Reflowing digital ink annotations. *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, pp. 385-93.

Bederson, B. B. (2004). Interfaces for staying in the flow. *Ubiquity*, 1-1.

Bonn, M. and McGlone, J. (2014). New Feature: Article Annotation with Hypothesis. *Journal of Electronic Publishing*, **17**(2).

Boyles, N. (2013). Closing in on Close Reading. *Educational Leadership*, **70**(4): 36–41.

Budd, J. W. (2004). Mind Maps as Classroom Exercises. *The Journal of Economic Education*, **35**(1): 35–46.

Buzan, T. and Buzan, B. (1993). The Mind Map Book How to Use Radiant Thinking to Maximise Your Brain's Untapped Potential. New York: Plume.

Geng, Z., Cheesman, T., Laramee, R. S., Flanagan, K. and Thiel, S. (2013). ShakerVis: Visual analysis of segment variation of German translations of Shakespeare's Othello. *Information Visualization*, **15**: 93-116.

Goffin, P., Willett, W., Fekete, J. D. and Isenberg, P. (2014). Exploring the placement and design of word-scale visualizations. Visualization and Computer Graphics, *IEEE Transactions*, **20**(12): 2291-300.

Hawthorn, J. (2000). *A glossary of contemporary literary theory*. Oxford University Press.

Jänicke, S., Blumenstein, J., Rücker, M., Zeckzer, D. and Scheuermann, G. (2015a). Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies. *Digital Humanities Quarterly*.

Jänicke, S., Franzini, G., Cheema, M. F. and Scheuermann, G. (2015b). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In Borgo, R., Ganovelli, F., and Viola, I. (eds.), *Eurographics Conference on Visualization (EuroVis) - STARs (2015)*, The Eurographics Association.

Kanter, B. (2015). Cambodia4kids.org, https://www.flickr.com/photos/cambodia4kidsorg/6195211411 (Retrieved 2015-11-25).

Kehoe, A. and Gee, M. (2013). eMargin: A Collaborative Textual Annotation Tool. *Ariadne*, **71**.

McCabe, M. M. (2015). *Platonic Conversations*. Oxford University Press.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Muralidharan, A. (2011). A Visual Interface for Exploring Language Use in Slave Narratives. *Proceedings of the Digital Humanities 2011*.

Schilit, B. N., Golovchinsky, G. and Price, M. N. (1998). Beyond paper: supporting active reading with free form digital ink annotations. *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co., pp. 249-56.

Sinclair, S. and Rockwell, G. (2012). Voyant Tools. Online: http://voyant-tools.org (Retrieved 2015-11-25).

Yoon, D., Chen, N. and Guimbretière, F. (2013). TextTearing: Opening white space for digital ink annotation. *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM, pp. 107-12.

Zyto, S., Karger, D., Ackerman, M. and Mahajan, S. (2012). Successful classroom deployment of a social document annotation system. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 1883-92.

Notes

[1] Image reproduced with permission from Kehoe (Kehoe et al., 2013)

[2] Image reproduced with permission from Kehoe (Kehoe et al., 2013)

[3] Image reproduced with permission from Kanter (Kanter, 2015) (Figure under CC BY 2.0 license, see https://creativecommons.org/licenses/by/2.0/ for details).

Modelling between Digital and Humanities: Thinking in Practice

Arianna Ciula
arianna.ciula@roehampton.ac.uk
University of Roehampton, United Kingdom

Øyvind Eide
oyvind.eide@uni-passau.de
University of Passau, Germany; University of Cologne, Germany

Cristina Marras
cristina.marras@cnr.it
ILIESI, CNR, Italy

Patrick Sahle
sahle@uni-koeln.de
University of Cologne, Germany

This poster will present the rationale behind as well as the work in progress of an international collaborative project funded by the Volkswagen Foundation (scheme *"Original – isn't it?"* New Options for the Humanities and Cultural Studies, Funding Line 2 Constellations), 2016-2017. The project aims to link scholarly modelling as a formal and informal reasoning strategy across disciplinary boundaries, and to bridge between modelling in research and teaching.

In Digital Humanities (DH), modelling is a creative process of reasoning in which meaning is made and negotiated through the creation and manipulation of external representations. Through the lenses of critical humanities traditions and interdisciplinary takes on making and using models, the project *Modelling between digital and humanities: thinking in practice* builds on the novelty of DH research in making explicit and integrating existing diverse models of cultural phenomena (e.g. texts; events) with the aim to:

- explore possibilities for a new interdisciplinary language of modelling;
- analyse modelling in scholarship as a process of signification;
- develop connections between modelling as research and learning strategies.

By modelling we intend the creative process by which researchers create and manipulate external representations ("imaginary concreta," Godfrey-Smith, 2009: 108) to make sense of the conceptual objects and phenomena they study. In order to integrate diverse theoretical frameworks around modelling (McCarty, 2005, 2009; Mahr, 2009; Frigg and Hartmann, 2012; Morgan, 2012; Kralemann and Lattmann, 2013; Flanders and Jannidis, 2015) with a practical dimension, the project makes use of DH as an interdisciplinary departure to study modelling as anchored both to computer science and to the humanities.

Research focus

Our working hypothesis is that in DH research, implicit and explicit models of cultural phenomena are integrated into external metamodels, e.g. graphical representations, which often embed natural language and are informal. These metamodels are iteratively translated towards computable implementations via a variety of more or less formal models: models for.

Two case studies are used to reflect on modelling in practical terms:

- Textuality, standing for the complexity of cultural objects and activities addressed by a plethora of subject-specific approaches. Sahle (2013) proposes a metamodel to chart and relate single models of textuality from several disciplines. The metamodel acts both as a **model of**the phenomenon of textuality and as a **model for** working with texts to inform the development of text technologies, digitisation practices, and rules for transcription and annotation.
- Events. While textuality mediates the world we live in, events are central to an epistemological perception and description of the processes shaping this world. Many disciplines contribute to theoretical reflections on and practical applications of the modelling of events (see, e.g., Le Boeuf et. al., 2015). The project aims at combining contributions from philosophy, literary studies, history, linguistics, and computer science with cultural heritage documentation and the news industry in the transition from **models of** events as things to perceive and talk about to **models for** event detection and description.

The analysis of modelling practices in the areas outlined above will aim at gaining new insights in the epistemology of modelling:

- How are theory and practice blended in these modelling efforts?
- What role do formal and informal metamodels play in translating models of cultural phenomena into implementations?
- What shared terminology can help us gaining an integrative and non-reductive understanding of digital modelling?
- Can we define the methods of digital modelling informed by such an integrative and non-reductive approach?

Societal resonance

The rationale of the project places the practice of DH within a broad understanding of how humans think through things. Models are ubiquitous in our contemporary society as powerful tools to schematise the complexities of our universe, from genes to climate, from the economy to the stars. By linking DH practices to the craft of computer science as well as to the critical humanities tradition, this project tackles issues at the centre of the construction and deconstruction of (digital) models. It also advocates for a critical DH research offering the instruments to unpack the rhetoric of digital and data models, so as to contribute to a pedagogy of the digital age and to act at the core of a new cultural literacy.

Outcomes

Over 18 months, the project aims at producing: 1) an open access book about modelling, and 2) An international workshop devoted to selected controversies around the theorisation and practice of modelling (e.g. fictions vs. non-fiction; theory vs. data), which will give important input to the book. A consulting group will be set up to discuss draft chapters and ongoing work.

Bibliography

Boeuf, P. Le, Doerr, M., Ore, C. E. and Stead, S. (2015). Definition of the CIDOC Conceptual Reference Model. Version 6.2. http://www.cidoc-crm.org/docs/cidoc_crm_version_6.2.pdf (accessed 18 February 2016).

Flanders, J. and Jannidis, F. (2015). *Knowledge Organization and Data Modeling in the Humanities*. White paper http://www.wwp.northeastern.edu/outreach/conference/kodm2012/ flanders_jannidis_datamodeling.pdf (accessed 18 February 2016).

Frigg, R. and Hartmann, S. (2012). Models in science (Ed.) Zalta, E. N. *The Stanford Encyclopedia of Philosophy*. Stanford University http://plato.stanford.edu/archives/fall2012/entries/models-science/ (accessed 18 February 2016).

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies*, **143**(1): 101–16.

Kralemann, B. and Lattmann, C. (2013). Models as icons: modeling models in the semiotic framework of Peirce's theory of signs. *Synthese*, **190**(16): 3397–420.

Mahr, B. (2009). Information science and the logic of models. *Software & Systems Modeling*, **8**(3): 365–83.

McCarty, W. (2005). *Humanities Computing*. Basingstoke [England]; New York: Palgrave Macmillan.

McCarty, W. (2009). Being reborn: the humanities, computing and styles of scientific reasoning. In Bowen, W. R. and Siemens, R. G. (eds), *New Technologies in Medieval and Renaissance Studies*, Tempe, Arizona: Iter and the ACMRS, vol. **1**: 1–23.

Morgan, M. S. (2012). *The World in the Model: How Economists Work and Think*. Cambridge University Press.

Sahle, P. (2013). Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung. [Preprint-Fassung] Universität zu Köln Ph.D. http://kups.ub.uni-koeln.de/5013/ (accessed 18 February 2016).

Using Digital Editions to Analyze Iliadic Text Reuse and its Poetic Tradition

Brian Robert Clark
brclar15@g.holycross.edu
College of the Holy Cross, United States of America; Center for Hellenic Studies, United States of America

Claude Spalding Hanley
cshanl18@g.holycross.edu
College of the Holy Cross, United States of America

Stephanie Clare Neville
scnevi17@g.holycross.edu
College of the Holy Cross, United States of America

Charles John Schufreider
cjschu17@g.holycross.edu
College of the Holy Cross, United States of America

Melody Anne Wauke
mawauk17@g.holycross.edu
College of the Holy Cross, United States of America

Since its foundation, the Homer Multitext project has been creating digital diplomatic editions of manuscripts of the *Iliad*, editions which record every intentional mark on the manuscript, in order to create a resource to explore larger Homeric questions. Our poster presents one aspect of Homer Multitext research. The authors' original research involves analyzing the Venetus A to look for multiforms, that is, substitutions for Homeric text that fit both metrically and contextually. The Venetus A, a tenth-century manuscript currently housed in Venice, is the oldest complete manuscript of the *Iliad* existing today. In addition to the intact poem, the Venetus A includes an abundance of scholia, scholarly notes which comment on the poem. It is in these scholia that we find the evidence for multiforms. Our use of the term "multiform" intentionally follows Albert Lord's of the *Iliad* as an orally composed and transmitted poem: "the word multiform is more accurate than variant, because it does not give preference or precedence to any one word or set of words to express an idea; instead it acknowledges that the idea may exist in several forms." (Lord, 2000). Ultimately, in looking for multiforms, we found a definitive correlation between the appearance of multiforms in the scholia and the reference to one of three Homeric textual editors, all of whom served as the head of the library at Alexandria at various times from the third to second centuries BCE. While it has been previously hypothesized that the work of these three Alexandrian editors, preserved in the Iliadic scholia, is responsible for the current standard canon of available multiforms, the research represents our first steps in using statistics to better understand the transition from an oral tradition to today's critical print editions.

The first task in our research was to analyze how many multiforms existed in a select portion of the *Iliad*. Books 18 and 19 of the *Iliad* served as our sample, and we accessed the manuscript through openly-licensed digital photography. From these photographs we created digital diplomatic editions of the poem and its corresponding scholia using an XML editor and following TEI guidelines. Within our editions we used the TEI element "quote" to identify every instance of text reuse found in the scholia, and then generated a separate table in which we classified the types of text reuse. Our classification was modelled on work by Monica Berti who defines "text reuse" as "the meaningful reiteration of text, usually beyond the simple repetition of common language" https://wiki.digitalclassicist.org/Text_Reuse). By assigning distinct URNs to the different types of text reuse in the table and not in our XML mark-up, we applied a unique approach that kept our external analysis separate from our digital edition.

We automatically generated a table displaying the occurrence of multiforms in relation to references to Alexandrian editors in preparation for a chi-square test for independence.

This table shows 34 instances of overlap, in which one of the scholars was named with a multiform. For the chi-square test, we compared these two sets of data, beginning with the null hypothesis that they were independent. Then, having set a significance level of 0.01, we obtained a p-value of $1.2 *10^{-42}$. This p-value shows that the probability of randomly obtaining the results we did is negligibly low. Thus, we were able to reject our null hypothesis and conclude that there is a correlation between the occurrence of a multiform and references to one of the Alexandrian grammarians.

Given the correlation between Alexandrian editors and multiforms attested to in the Venetus A, we conclude that the Alexandrian editors were responsible for setting the canon of multiforms which has been transmitted in this manuscript, and from there to the modern world. In respect to further avenues of research, the strength of this correlation remains to be tested. Additionally, we have not yet attempted text reuse analysis with the majority of the Venetus A; drastically increasing the sample size will likely yield more insights into the nature of the poem's transmission and production. The ability to statistically support this conclusion exemplifies the great promise statistical, analytical methods hold for the field of Homeric studies, as well as wider areas in other fields of the humanities.

Bibliography

Lord, A. (2000). *Singer of Tales (2nd ed.)*. Cambridge, MA.

Notes

[1] The following is the link to a table which shows our classification of the instances of text reuse. (https://github.com/hmteditors/hc-il18/blob/master/venA/collections/reuse_18_19.csv)

[2] The following is the link to data tables showing the correlating frequencies of appearances of multiforms and Alexandrian editors. (https://github.com/hmteditors/hc-il18/blob/master/venA/stats/Table%20of%20multiforms.md)

Editing in a text-image-sound form: the eTalks

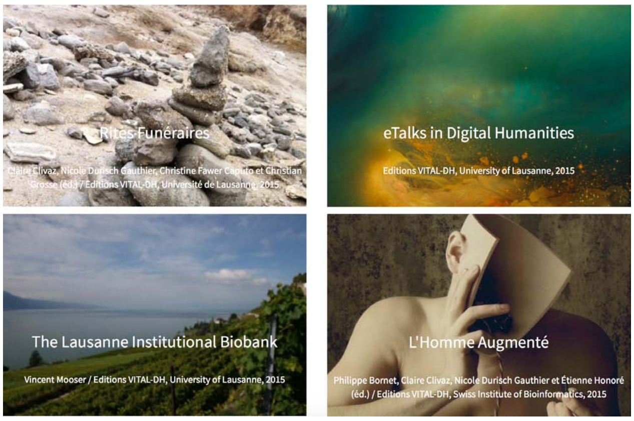
Claire Clivaz
claire.clivaz@isb-sib.ch
Swiss Institute of Bioinformatics, Switzerland

Martial Sankar
martial.sankar@isb-sib.ch
Swiss Institute of Bioinformatics, Switzerland

Cecile Pache
cecile.pache@unil.ch
University of Lausanne, Switzerland

An interdisciplinary team of researchers has built a new multimedia editorial tool: the eTalks (Clivaz 2014; Clivaz et al., 2015a; Clivaz et al., 2015b; EADH projects, 2016), based primarily on speeches of scholars. Simple videos or MP3 recordings of lectures may prove insufficient to many researchers since they are unquotable in detail and they do not offer the possibility of being combined with text, images, hyperlinks, and references. Until now, no tool has been available for creating a carefully edited product that includes text-image-sound, all entirely quotable in details: yet, this is what we have achieved with the eTalks [http://etalk.vital-it.ch/mooser/mode-demploi-en/].

In creating the eTalks, we were motivated by the fact that academic publications and pedagogy have been deeply reconfigured by the emergence of a new kind of knowledge produced by the synergy between text, image and sound. As Tanya Clement points out, diverse Digital Humanities (DH) pedagogies, such as new media studies and game studies, can be characterized by looking at multiliteracies "that are engaged within undergraduate humanities curricula through general skills, principles and habits of mind that allow students to progress within and engage society in the twenty-first century" (Clement 2012). Academic publications in Humanities are slower than pedagogy in terms of the testing of multimodal literacies. However, different tools are now able to present slides joined to videos of scholarly talks, such as Slideshot and Dashboarding [http://slideshot.epfl.ch/play/cops_binney; http://www.infoq.com/presentations/dashboard-data-analysis?utm_source=infoq&utm_medium=related_content_link&utm_campaign=relatedContent_presen], but they cannot be quoted in detail. Scalar, a very impressive multimodal tool, proposes to "create interpretive pathways through the materials"[http://scalar.usc.edu/about/], privileging users' points of view. The eTalks claim to rely on the scholar's oral talk as a leading way among multimodal materials while giving the users the possibility of reconsidering the auctorial point of view by directly accessing all the quoted sources.


Rites Funéraires
Claire Clivaz, Nicole Durisch Gauthier, Christine Fawer Caputo et Christian Grosse (éd.) / Editions VITAL-DH, Université de Lausanne, 2015
eTalks in Digital Humanities
Editions VITAL-DH, University of Lausanne, 2015
The Lausanne Institutional Biobank
Vincent Mooser / Editions VITAL-DH, University of Lausanne, 2015
L'Homme Augmenté
Philippe Bornet, Claire Clivaz, Nicole Durisch Gauthier et Étienne Honoré (éd.) / Editions VITAL-DH, Swiss Institute of Bioinformatics, 2015

The eTalks application implements an easy-to-use editor interface, designed for the use of researchers themselves, allowing for the creation and editing of original enhanced talks. This permits the linking together of images, sounds and textual materials by means of hyperlinks, thereby enriching the content with relevant information. The result of the editing is displayed through a viewer interface, allowing one to experiment with the entire eTalk or to actively navigate, scroll and search inside its content. After recording the speech of the scholar, the Audacity software allows for the splitting of the speech into pieces of 2-3 sentences. Each piece of speech can be associated with its written version, a slide, images, or hyperlinks and so forth. Each piece is also quotable with a specific URL: a new kind of reference. Thus, the final release of eTalks allows for the complete 'citability' of its contents: each and every portion of the researchers' talks can be precisely referred to and therefore cited, just like any traditional, paper-based scientific publication but with all the potential for plural literacies.

The core of the eTalk engine was developed in JavaScript and the code is now available as open source on Github as a free application for further development. The eTalks are currently being further developed and disseminated by an interdisciplinary team of researchers in Digital Humanities and bioinformatics at the Swiss Institute of Bioinformatics (Lausanne). Four series of eTalks have thus far been published as openly accessible: twelve on funerary rituals, nine on the enhanced Human, two on the institutional biobank of Lausanne, and one in Digital Humanities [http://etalk.vital-it.ch]. The eTalks are now in development by institutional and research collaborations, notably the Pedagogical High School of Lausanne (HEPVaud) and the ERASMUS+ #dariahTeach project, whose purpose is to offer a webportal by 2017 that will include digital teaching modules [www.dariah.eu/teach].

We will present eTalks' main features in our poster, and in particular the question of copyrights: to be able to quote several images, the team had to learn the basic rules of the relevant Swiss laws, and how Wikipedia commons work in Switzerland; furthermore we plan to develop European test-cases. We have also learned to negotiate with the authors and to convince them to rather use open access material. In difficult cases, we consult specialized people. Such obstacles had to be navigated and new skills acquired by our team as new, necessary knowledge. We will secure the interoperability of our data, and progressively introduce videos, and purl references.

Bibliography

Clement, T. (2012). Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind. In Hirsch, B. (ed.) *Digital Humanities Pedagogy: Practices, Principles and Politics*, Cambridge, UK: Open Book Publishers, pp. 365–88. http://www.openbookpublishers.com/htmlreader/DHP/chap15.html

Clivaz, C. (2014). De l'article à l'etalk : enjeux et défis de la littératie plurielle dans la communication académique, *Actes du colloque de l'AIPU 2014, Mons (Belgique)*. http://hosting.umons.ac.be/php/aipu2014/C9TEST/select_depot2.php?q=1775

Clivaz, C., Rivoal, M. and Sankar, M. (2015a). A New Platform for Editing Digital Multimedia: The eTalks. In Schmidt, B. and Dobreva, M. (Eds.) *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science*, The authors and IOS Press, doi: 10.3233/978-1-61499-562-3-156; http://ebooks.iospress.nl/publication/40894

Clivaz, C., Pache, C., Rivoal, M. and Sankar, M. (2015b). Multimodal literacies and academic publishing: the eTalks, *Information Services and Use*, **35**: 4.

EADH projects (2016). http://eadh.org/projects/etalks

Discursive Constructions Of Culture: A Semantic Model For Historical Travel Guides.

Ulrike Czeitschner
Ulrike.Czeitschner@oeaw.ac.at
Austrian Academy of Sciences, Austria

Context

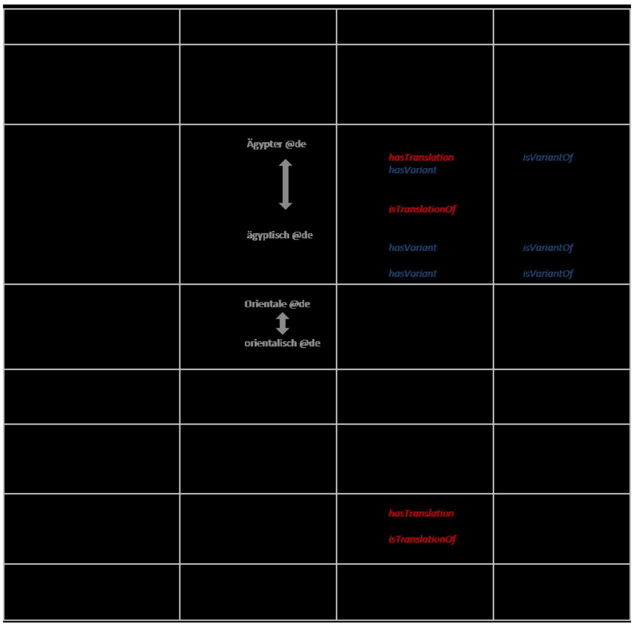
The *Baedeker Corpus*, a digital collection of early German travel guides that is currently being developed in the *travel!digital* project[1], brings together a valuable but rarely investigated part of cultural heritage in an up-to-date and sustainable digital form. A combination of source oriented approaches and Semantic Web formats (Meroño-Peñuela et al, 2014) allows for enhanced access to both the TEI-based digital edition and the rich domain-specific knowledge. As complex inter-texts (Wierlacher, 1997; Koshar, 2000) and significant discourse-historical artefacts (Maingueneau, 2014) travel guides represent codified and authorized versions of local culture and history (Pritchard et al, 2005). Reflecting dominant discourses, (re)producing and (re)constructing them, the genre plays a central role in shaping the tourist experience and directing the tourist gaze (Thurlow et al, 2007; Urry, 1990). Giving insight into various readings of history, tradition and culture, the new language resource is meant to foster cross-disciplinary research in cultural representation and identity constructing discourses.

The *Baedeker Corpus* comprises *all* first editions of German travel guides on non-European countries which were brought out by the Baedeker publishing house before World War I. It contains more than 1.5 million running words and covers various regions, offering a balanced picture of different cultural areas.[2] Along with the basic layers of linguistic annotation, controlled vocabularies and content contextualization by *Linked Open Data* resources will assist in exploring cultural narratives from the turn of the 19th century.

bdk:ConceptScheme(s)

Focusing on *people* and *monuments,* two essential components of the guidebook genre and of cultural discourse itself, the extensive lexical inventory is represented by means of the *Simple Knowledge Organization System SKOS*. Besides personal names, people are addressed in a variety of different forms in the guidebooks. References to classes are frequently used in making generalizations about groups and individuals as well (cf. Schmidt-Brücken, 2015). Each of these generic expressions is represented by a *bdk:Descriptor*, which is defined as a subclass of *skos:Concept*. Individual terms are encoded as *skosxl:prefLabel(s)* and *skosxl:altLabel(s)*. Variants and translations are designated on term-level by the properties *hasVariant/isVariantOf* and *hasTranslation/isTranslationOf*. Modelled in this way, the vocabulary identified in the guidebooks forms the basis of the concept scheme. Seven categories indicated by *skos:topConceptOf* improve structuring of the bdk:ConceptSchemeGroups:

1. Collective terms: *population, tribes, natives;* **2.** ethnic/national communities: *Englishmen, Wedda;* **3.** geographical concepts: *Europeans, Orientals;* **4.** names of languages and scripts, language affiliation: *Arabic, Cyrillic;* **5.** professions, including political, religious, economic roles, and styles of living: *merchants, government officials, priests, peasants, nomads;* **6.** religious communities: *brotherhood, pilgrims; Buddhists, Sikhs;* and **7.** social classes: *castes, workers.* Concepts and labels include both nouns and adjectives, indicating associations among them by means of the property *skos:related*. Figure 1 lists definitions of *skos:topConcept(s)* and shows selected examples of concepts and labels.



	Ägypter @de ↕ ägyptisch @de	*hasTranslation* *hasVariant* *isTranslationOf* *hasVariant* *hasVariant*	*isVariantOf* *isVariantOf* *isVariantOf*
	Orientale @de ↕ orientalisch @de		
		hasTranslation *isTranslationOf*	

Figure 1: bdk:ConceptSchemeGroups.

The picture is similarly varied for monuments and notable sights. Since assessments and classifications of cultural heritage objects are integral parts of cultural representation, they are included in a separate concept scheme. The bdk:ConceptSchemeMonuments is organized by skos:topConcept(s) which indicate topical spheres the objects belong to, ranging from architecture and artworks, to accommodations, landscapes, and breath-taking views.

Prospects

By the end of the project, a web application based on the *corpus_shell* framework[3] will make the digital texts available along with their facsimiles, exposing them via FCS/

SRU protocol[4] that is part of the CLARIN infrastructure. This online edition will offer querying capabilities inside of both the text and the linguistic annotation layers. It will provide indexes of word classes, lemmas and the semantic entities defined by the SKOS-vocabularies presented in this abstract. Transforming names of *people/s* and *monuments* in the texts into links to the LOD cloud[5], the vocabularies will connect occurrences in the *Baedeker Corpus* to other online resources, providing enhanced access to the corpus and additional information via the guidebooks' main actors. The presented data model aims at supporting fine-grained examinations of semantic components that have a lasting influence on cultural perceptions of "Other" and "Self". It is expected that semantic technologies do have the potential to reveal much about a discourse that goes far beyond travel literature.

Bibliography

Clarin E. (n.d.). Federated Content Search (CLARIN-FCS). https://www.clarin.eu/content/federated-content-search-clarin-fcs (accessed 22 December 2015).

Ďurčo, M. and Mörth, K., et al. (n.d.). corpus-shell. https://clarin.oeaw.ac.at/corpus_shell (accessed 22 December 2015).

Jaworski, A. and Pritchard, A. (eds.) (2005). *Discourse, communication and tourism.* Clevedon: Channel View Press.

Koshar, R. (2000). *German Travel Cultures.* Oxford: Berg.

Maingueneau, D. (2014). Diskurs und Äußerungsszene. Zur gattungsspezifischen Kontextualisierung eines Zeitungsartikels zum unternehmerischen Bildungsdiskurs. In Angermuller, J. and Nonhoff, M. et al. (eds.) *Diskursforschung. Ein interdisziplinäres Handbuch* (2 Bde.). Bielefeld: transcript Verlag, pp. 433-53.

Meroño-Peñuela, A. and Ashkpour, A., et al. (2014). Semantic Technologies for Historical Research: A Survey. *Semantic Web*, IOS Press, pp. 1-27. http://www.semantic-web-journal.net/sites/default/files/swj301.pdf (accessed 22 December 2015).

Schmidt-Brücken, D. (2015). *Verallgemeinerung im Diskurs. Generische Wissensindizierung in kolonialem Sprachgebrauch.* Berlin/München/Boston: Walter de Gruyter GmbH.

Stehouwer, H. and Durco, M., et al. (2012). Federated Search: Towards a Common Search Infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. ELRA, pp. 3255–59.

Urry, J. (2002 [1990]). *The Tourist Gaze. Leisure and Travel in Contemporary Societies.* London: SAGE Publications.

Wierlacher, A. (1997). Verfehlte Alterität. Zum Diskurs deutschsprachiger Reiseführer über fremde Speisen. In Teuteberg, H. J. and Neumann, G. et al. (Hrsg.). *Essen und kulturelle Identität. Europäische Perspektiven.* Berlin: Akademie Verlag GmbH, pp. 498-509.

Notes

[1] The project "travel!digital. Exploring *People* and *Monuments* in Baedeker Guidebooks (1875–1914)" is funded by the platform *Digital Humanities Austria*.

[2] Palestine and Syria (1875), Lower (1877) and Upper Egypt (1891), North America and Mexico (1893), Asia Minor (1905), the Mediterranean Coastline of Africa (1909), India and Ceylon (1914).

[3] Cf. https://clarin.oeaw.ac.at/corpus_shell [accessed 2015-12-22].

[4] Cf. https://www.clarin.eu/content/federated-content-search-clarin-fcs [accessed 2015-12-22].

[5] LOD datasets: VIAF Virtual International Authority File, GESIS Thesaurus for the Social Sciences, AAT Art & Architecture Thesaurus, UNESCO Thesaurus, DBpedia.

Visualizing the Gradual Production of a Text

Peter Daengeli
pdaengeli@me.com
Cologne Center for eHumanities (CCeH)

Christian Theisen
ctheise1@smail.uni-koeln.de
Cologne Center for eHumanities (CCeH)

Magnus Wieland
magnus.wieland@nb.admin.ch
Swiss Literary Archives (SLA)

Simon Zumsteg
simon.zumsteg@nb.admin.ch
Swiss Literary Archives (SLA)

Introduction

The *Lokalbericht* project (2013-2016)[1] aims to deliver a scholarly digital edition of Swiss author Hermann Burger's first novel *Lokalbericht*. Accompanied by a reading edition in print (fall 2016), the digital edition will include all extant precursory stages of the novel as well as Burger's final revision of the text, written between 1970 and 1972 (Wieland et al., 2012). The project relies on document oriented encoding, as supported by the TEI guidelines since P5 v. 2.0 (cf. Burnard et al., 2012), in order to aptly capture the materiality and the genetic dimension of the work.

Macrogenetic Challenges

Whereas the recommendations of other textgenetic edition projects as well as the tradition of genetic editions in print were very helpful for the encoding of local, microgenetic phenomenons (i.e. intradocumental) and their presentation to the reader, much less information was available on how to document interdocumental textgenetic relations in a structured manner. In order to better understand the challenges and needs of what we dubbed macrogenetic editing, a workshop was convened in Bern in 2014.[2] As Hans Walter Gabler concluded, 'a digital genetic edition will only tap its full potential by capturing and visualizing "macrogenetic" aspects' (Gabler, 2014).

Incited by the outcomes of the workshop we investigated ways that offer immediate insight into Burger's way of writing – the novel was crafted as a mosaic consisting of numerous tesserae that were worked over repeatedly. This writing technique results in an f *avant texte* consisting of many segmented and related texts, of which at times more than a dozen copies are extant. Taking a bird's eye view of the corpus, the reader should get a clear understanding of the writing process and be able to follow the development of specific segments or chapters and their integration into the final typescript, before swooping down to the text level to discover the textgenetic minutiae.

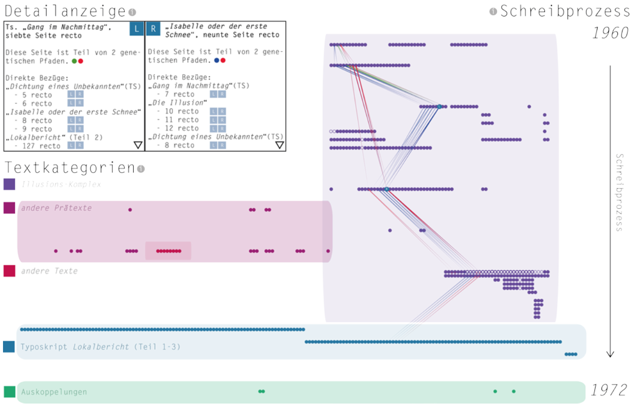
A Visual Approach to Macrogenetic Relationships

In printed and early digital genetic editions, macrogenetic relations are often visualized as a stemma. Such stemmata leave much to be desired – they often abstract on a high level (e.g. witnesses or versions of a text) as opposed to smaller textual entities such as paragraphs or sentences, require considerable intellectual and manual effort, and remain comparatively static, perhaps linking to representations of related documents, but lacking in precision. Yet their genealogical tree structure serves still as a powerful basis to represent editorial changes.

Taking the underlying idea to represent temporal development on the vertical axis and introducing finer levels of granularity in a stemma tree on one hand threatens to lead to over-complexity and problems of positioning. Resorting to force directed plotting as it is often used to display and investigate graph data on the other hand does not satisfy the sequentiality inherent in written text. For the manageable amount of typescript pages and drafts of the novel at hand, we found that a prearranged layout that entails a temporal dimension (y-axis) and material evidence of the texts (x-axis) works best (cf. figure 1).[3] Genetic relations between the typescript pages, the smallest units of display,[4] are highlighted on mouse actions. Interacting with this visualization, the user of the digital edition can trace the development of the early and final drafts of the novel.

Accessing Data through Visual Navigation

Good textgenetic visualizations should not be mere by-products of digital editions, but closely tie in with the structure of the digital edition, i.e. serve as a subsidiary navigation to its contents. By way of two panels that yield lists of related typescript pages the user can access detail views in the main interface of the edition (single or synoptical views) directly from the visualization. Vice versa, the visualization initially highlights the current selection of the main interface in order to provide specific views of the *dossier génétique*.



Ts. „*Gang im Nachmittag*", siebte Seite recto

Diese Seite ist Teil von 2 genetischen Pfaden. ●●

Direkte Bezüge:
„*Dichtung eines Unbekannten*"(TS)
- 5 recto
- 6 recto
„*Isabelle oder der erste Schnee*"
- 8 recto
- 9 recto
„*Lokalbericht*" (Teil 2)
- 127 recto

„*Isabelle oder der erste Schnee*", neunte Seite recto

Diese Seite ist Teil von 2 genetischen Pfaden. ●●

Direkte Bezüge:
„*Gang im Nachmittag*"(TS)
- 7 recto
„*Die Illusion*"
- 10 recto
- 11 recto
- 12 recto
„*Dichtung eines Unbekannten*"(TS)
- 8 recto

Figure 1: Display of two genetically related sheets: genetic relations (r.) and navigational component (top l.)

Outlook

While the *analytical approach* based on genetic paths defined by the editors (on sheet level) should provide the most valuable knowledge about the production of the novel, it would be desirable to complement the visualization by an *explorative mode* based on relations encoded on more granular levels.[5] A visualization of this kind would be less useful to pursue particular authorial changes, but it would help to identify patterns that deserve closer attention and facilitate the isolation and display of the least interpretation-ridden relations and consequently better suit a reader-centered concept of a digital edition.

The wish for visualizations that abstract from the detailed genesis of a text to a more approachable overview is shared by a number of genetic edition projects.[6] After all, imparting his or her knowledge and insight to the readership lies at the core of a scholarly editor's role. We hope to contribute to the debate by sharing our approach to this problem at DH 2016 and finally by launching the digital *Lokalbericht* edition in October 2016.

Bibliography

Burnard, L., Jannidis, F., Pierazzo, E., Rehbein, M. (2012). An Encoding Model for Genetic Editions. http://www.tei-c.org/Activities/Council/Working/tcw19.html (accessed 1 March 2016).

Gabler, H. W. (2014). Personal report. In Schweizerisches Literaturarchiv, *Resümees zum internationalen Workshop* Digitale genetische Editionen (in der Praxis) vom 4./5. September 2014 im Schweizerischen Literaturarchiv (SLA), **6**. http://lokalbericht.unibe.ch/hermann_burger/pdf/Resuemees.pdf (accessed 1 March 2016).

Wieland, M. and Zumsteg, S. (2012). Lokalbericht. Von der Archivfiktion zur Archivedition. *Germanistik in der Schweiz*, **9**: 91-109.

Notes

[1] Cf. the preliminary project website at http://lokalbericht.unibe.ch.

[2] Cf. http://lokalbericht.unibe.ch/hermann_burger/workshops.html

[3] The actual technical implementation makes use of the D3.js JavaScript library (cf. http://d3js.org) for manipulating documents based on data, but similar outcomes might be realized using other tools.

[4] More complex genetic corpora and different choices of granularity would however require more sophisticated zooming and panning mechanisms or ways of three-dimensional stacking and it remains to be seen whether the chosen approach would serve them well.

[5] The development of an explorative visualization of the entire *Lokalbericht* corpus as initially planned will presumably only be implemented for selected texts due to limited resources.

[6] Examples are the digital historical critical edition of Arthur Schnitzler's works (1905–1931), the digital edition of Goethe's Faust, or the Samuel Beckett Digital Manuscript Project.

Bridging the Gap: the Digital Innovation Group

Julia Luise Damerow
jdamerow@asu.edu
Arizona State University, United States of America

Erick Bruce Peirson
bpeirson@asu.edu
Arizona State University, United States of America

Manfred Laubichler
manfred.laubichler@asu.edu
Arizona State University, United States of America; Santa Fe Institute; KLI, Klosterneuburg, Austria

The success of the digital humanities as a field of research has led to two pressing needs at the interface of computer science and the humanities. First, sustainable software development projects are needed that are focused on humanities research problems. Second, graduate and undergraduate training models are required that address the interdisciplinary nature of digital and computational humanities research (Ramsay, 2012; Reid, 2012). From the application and development of new algorithms to mine texts (e.g. Murdock et al., 2015) to the development of multi-institutional software systems (e.g. Neuroth et al., 2011), many digital humanities projects can only be implemented with the help of computer scientists and software engineers.

We believe that the future success of applying digital and computational approaches to research questions in the humanities depends on forging links between computer science and the humanities. However, the digital humanities need to involve more than adding a computer scientist to a humanities project. Many digital humanities programs include computer science courses in their degree and certificate programs, but this is not yet the norm. Specifically in the history and philosophy of science, many programs do not impart the skills necessary to take advantage of the new modes of research. Likewise, computer scientists typically receive little training in the humanities, making it difficult to foster meaningful and productive collaborations with humanities scholars.

Our poster describes the Digital Innovation Group (DigInG) in the Center for Biology and Society at Arizona State University (ASU). DigInG's primary objective is the development of computational solutions for questions in the digital humanities with a focus on digital history and philosophy of science. Our group trains students from the computer science department together with students from the humanities. We develop user-oriented innovative tools, methods, and infrastructures, and foster an understanding for each other's fields by teaching a skill set that enables students to communicate more effectively.

As part of DigInG, we developed a course that puts computer science students in a room with humanities students to work together on digital humanities projects. In addition, we offer the opportunity to work on the software development projects in our lab as student workers. We emphasize software engineering best practices and cross-disciplinary interaction. For example, in the beginning of the semester the class develops a course website, which requires them to learn how to use Git and GitHub. Students from the humanities learn the basics of programming. Computer science students acquire a mindset that better prepares them for working on digital humanities projects and learn how to better communicate with those outside their field. We discuss topics such as the importance of authority control, metadata, or resource discovery. In our experience, the students gain a lot from participating in our class as for many of them it is the first time that they work with students from a very different field of study. For instance, one humanities student noted that "learning to formulate questions in ways that computer scientists and software developers can understand" was one of the most valuable outcomes for them.

Not only does DigInG play an important role in developing software for humanities research, it also creates an environment for hands-on training for graduate and undergraduate students in computer science and the humanities. We hope that our poster will contribute to broadening the discussion about how digital and computational humanities programs are organized. We believe that software development and training in the digital humanities do not need to be separate endeavors.

Bibliography

Murdock J., Allen C., DeDeo S. (2015). Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. *CoRR*. http://arxiv.org/abs/1509.07175
(accessed 31 October 2015).

Neuroth H., Lohmeier F., Smith, K. (2011). TextGrid – Virtual Research Environment for the Humanities. *International Journal of Digital Curation*, **6**(2).

Ramsay, S. (2012). Programming with Humanists: Reflections on Raising an Army of Hacker-Scholars in the Digital Humanities. In Hirsch B. D. (ed.), *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge, UK: Open Book Publishers, pp. 227-39.

Reid, A. (2012). Graduate Education and the Ethics of the Digital Humanities. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 350-67.

The Attribution of the Lazarillo de Tormes. Shedding Some Light into a Centuries Old Problem

Javier de la Rosa Pérez
jdelaro@uwo.ca
The CulturePlex Lab, University of Western Ontario, Canada

Summit work of the Spanish Golden Age and forefather of the so called picaresque novel, The Life of Lazarillo de Tormes and of His Fortunes and Adversities (henceforth: the Lazarillo still remains an anonymous text (Rico, 2011). The 400 years of attributions have left us an enormous, nearly intractable, amount of bibliography that must be reviewed and studied. Paradoxically, scholars, instead of shying away from this mystery, are still adding new proposals to the pool of candidate authors, although some of them use modern and less explored methods (mostly computational) that were not available a decade ago.

Chronologically, the Hieronymite Friar José de Sigüenza was the first to propose a possible author: Friar Juan de Ortega. Father Sigüenza's Historia de la Orden de San Jerónimo gathers his finding of a manuscript of the Lazarillo in the cell of Juan de Ortega. Although a draft was indeed found in the Friar's cell, the circulation of handwritten copies was a common practice during the Spanish Golden Age (Botrel and Salaün, 1974). The claim that Father Ortega was the author is hard to sustain as the draft does not seem to be enough proof. Diego Hurtado de Mendoza was proposed a couple of years later, in 1607. His candidacy as the author of the Lazarillo was proposed by Valerio Andrés Taxandro. Over the years other scholars contributed to the diffusion of Mendoza being the author, and the attribution proved to be extremely popular. For almost three centuries book catalogues all over Europe recorded Hurtado de Mendoza as the author of theLazarillo. In 2010 Mercedes Agulló provided documentary proof to support the authorship, although some dispute the validity of such evidence (Agulló y Cobo, 2011). Other humanists were also proposed. The reformist Juan de Valdés was defended by Morel-Fatio and Manuel J. Asensio, but in view of the lack of solid evidence, this candidacy was abandoned in favor of his brother Alfonso. Rosa Navarro Durán supported Alfonso de Valdés, as she found connections between the works that influenced both Valdés' work and the Lazarillo. The lack of direct comparisons has been strongly criticized.

In the last decade, a couple of names have taken center stage. Juan Luis Vives was proposed and devotedly defended by Francisco Calero. Following the same precepts as for Alfonso de Valdés' candidacy, the fact that all Vives' known works were written in Latin makes the attribution lose some credibility. In 2008, after abandoning the authorship of Francisco Cervantes de Salazar, José Luis Madrigal proposed Juan Arce de Otálora. Using existing corpora such as Google Books and CORDE, Madrigal employed basic computational analysis based on the counting of words in order to find correlations between the style of authors in the corpora and the style of the author of the Lazarillo.

Computational approaches to authorship attribution are usually not considered to be enough proof to state the final truth in the disputed authorship of an anonymous text. However, the use of modern authorship attribution techniques might help to reduce the pool of candidates and contribute with evidence to support a specific possible author. After building the pool of candidates, we collected each candidate's available texts. Half of them lacked modern editions, which we solved building and using our own crowdsourcing OCR reviewing tool. The corpus we created counts a total of 50 works in a 90 year period surrounding the publication of the first known edition of the Lazarillo in 1554. All the major candidates for the authorship of this book were included, as well as some authors who had not been considered previously to add robustness to our analysis. The rules we followed for regularizing the spelling of old Spanish were borrowed from Ocasar's system (Ocasar, 2014). Features from the text were then extracted following the winners of several editions of the authorship attribution competition known as PAN at CLEF, which establishes the state-of-the-art in attribution techniques (Stamatatos et al., 2015). The final set of features were composed by several distributions: functions words, the 300 most common words (BOW), the 3000 most common character 3-grams (CNG), punctuation signs, the 30 most common parts of speech; tf-idf for a maximum of 1000 word bi- and tri-grams, and for a maximum of 1000 character {2,4}-grams; and average sentence length, sentence length variation, and sentence lexical diversity. A combined feature vector with all the abovementioned features was also included.

Once the texts were in digital format, we explored the dataset through distance-based methods, such as Burrows' Delta and its variations (Burrows, 2003), which outperformed any other. For compression-based methods we applied Cilibrasi and Vitanyi's NCD with BZIP2, RAR and PPM (Cilibrasi and Vitanyi, 2005). We tested PCA and Linear Discriminant Analysis with different settings for number of chunks per work, components and features (Burrows and Hassall, 2002). Our final approach was comprised of three steps: first we used unsupervised learning to reduce the pool of candidates. Then, applying supervised learning, we ranked the possible authors. Finally, only six of these candidates were fed into an ensemble algorithm for "unmasking" the most likely author. As for unsupervised learning techniques, we obtained that Ridge, Bernoulli, multinomial, and nearest centroid had the best performance for our total feature vector, BOW, and CNG. In supervised learning the best results were provided by SVM and maximum entropy models, for the same feature sets. This allowed us to reduce the pool of candidates for the unmasking method proposed by Moshe Koppel and Jonathan Schler since it is a computationally expensive method (Koppel and Schler, 2004). The results were consistent for all methods and in line to what we first obtained applying Burrows' Delta. We found that the most likely author seems to be Juan Arce de Otálora, closely followed by Alfonso de Valdés. Unfortunately, although supporting previous hypotheses about the authorship of the Lazarillo and providing with evidence in the case of Valdés, the method also stated that no certain attribution could be made with the given corpus.

Bibliography

Agulló y Cobo, M. (2011). A vueltas con el autor del Lazarillo. Un par de vueltas más . *Lemir: Revista De Literatura Española Medieval Y Del Renacimiento,* **15**: 217-34.

Anónimo (2011). Lazarillo de Tormes. In F. Rico (ed.) *Madrid: Real Academia Española-Galaxia Gutenberg-Círculo de Lectores* .

Asensio, M. J. (1959). La intención religiosa del Lazarillo de Tormes y Juan de Valdés. *Hispanic Review*, pp. 78-102, ISSN 0018-2176.

Botrel, J. and Salaün, S. (1974). *Creación y público en la literatura española.* Editorial Castalia.

Burrows, J. (2003). Questions of Authorship: Attribution and Beyond. *Computers and the Humanities*, **37**(1): 5.

Burrows, J. F. and Hassall, A. J. (1988). Anna Boleyn and the authenticity of Fielding's feminine narratives. *Eighteenth Century Studies*, pp. 427-53.

Cilibrasi, R. and Vitanyi, P. (2005). Clustering by compression. *Information Theory, IEEE Transactions On*, **51**(4): 1523-45.

De la Concha, V. (1972). La intención religiosa del Lazarillo. *Revista De Filología Española*, **55**(3): 243-77, ISSN 1988-8538.

Koppel, M. and Schler, J. (2004). Authorship verification as a one-class classification problem. *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Ocasar, J. L. (2014). La atribución del Lazarillo a Arce de Otálora. Una perspectiva geneticista sobre los estudios de autoría, 2014.

Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P. and López-López, A. (2015). Overview of the Author Identification Task at PAN 2015. *CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*, **1391**: 31, ISSN 1613-0073.

Defining the Core Entities of an Environment for Textual Processing in Literary Computing

Angelo Mario Del Grosso
angelo.delgrosso@ilc.cnr.it
National Research Council, Italy

Davide Albanesi
davide.albanesi@ilc.cnr.it
National Research Council, Italy

Emiliano Giovannetti
emiliano.giovannetti@ilc.cnr.it
National Research Council, Italy

Simone Marchi
simone.marchi@ilc.cnr.it
National Research Council, Italy

Introduction

The development of applications in the field of Digital Humanities (DH) does not adequately take into account domain modelling, software design principles and software engineering methodologies (Bozzi, 2013; D'Iorio, 2015; McCarty, 2008; Terras and Crane, 2010). In fact, many systems developed in the context of DH-related projects have not been conceived to be modular, extensible, and scalable: they only tend to solve specific problems such as data-driven and project-oriented tools (Boschetti and Del Grosso, 2015). In addition, most projects focus on the requirements of humanists (as end users), but leave out the needs of software developers.

This research was motivated by a number of issues emerged from the projects we worked on (Abrate et al., 2014a; Albanesi et al., 2015; Bellandi et al., 2014; Bozzi, 2015; Del Grosso, 2013; Ruimy et al., 2012) and it fits into an ongoing discussion about textual modelling and research infrastructures (Moulin et al., 2011; Pierazzo, 2015;

Schmidt, 2014). In particular, this work aims at providing methodological guidelines for the definition of the core entities of a digital scholarly environment. We chose to adopt an object-oriented approach since it can bring benefits in the definition of efficient and effective digital tools (Boschetti et al., 2014; Del Grosso and Nahli, 2014). To give an analogy, the environment we propose can help developers and scholars as CMS (e.g. Wordpress) can help Web designers and publishers.

The development of the environment follows three criteria: i) adopting an agile process (Ashmore and Runyan, 2014) to define the nature and behavior of the environment through both functional (e.g. user stories) and non-functional requirements (e.g. data model, system architecture) (Cohn, 2004; Collins-Cope et al., 2005); ii) providing well-defined Application Programming Interfaces (APIs) among components (Grill et al., 2012; Tulach, 2008); iii) applying analysis, architectural and design patterns for the sake of abstraction, generalization and flexibility (Ackerman and Gonzalez, 2011; Buschmann et al., 2007; Gamma et al., 1995).

Following the agile methodology, we are developing a modular environment by starting from the design and implementation of a **microkernel** (Buschmann et al., 1996) as the manager of the different components. In addition, the microkernel provides all the operations needed to manipulate the domain basic entities which are described in the section "Domain entities and design patterns".

Related works

Digital humanists have access to several tools for literary studies. TextGrid, for example, provides integrated tools for analyzing texts and gives computer support for digital editing purposes (Hedges et al., 2013). The NINES project offers an environment to support scholars in the creation of long-term digital research materials. It includes three main tools: Collex (Nowviskie, 2007), Juxta, and Ivanhoe. Annotation Studio is a collaborative system to annotate texts and add links to multimedia objects (Paradis et al., 2013). The CULTURA project aims at developing a "corpus agnostic research environment" providing customizable services for a wide range of users (Steiner et al., 2014). The development of an online workspace which helps scholars in the production of critical editions is the main objective of the Workspace for Collaborative Editing framework (Houghton et al., 2014). It uses existing standards and open-source solutions to create an architecture of reusable components. Other platforms worth mentioning are TUSTEP/TXSTEP (Ott, 2000; Ott and Ott, 2014), WebLicht (Hinrichs et al., 2010), Perseids (Almas and Beaulieu, 2013), Muruca/Pundit (Grassi et al., 2013), Textual Communities (Bordalejo and Robinson, 2015), SAWS (Jordanous et al., 2012), Voyant Tools (Sinclair and Rockwell, 2012), Transcribe Bentham (Causer and Terras, 2014) and Alcide (Moretti et al., 2014). However, the aforementioned initiatives allow digital scholars to meet specific needs, but none of them seems to provide, simultaneously, all the following characteristics: i) reusability and extensibility, ii) ease of use and configuration, iii) continuous availability of the services and development over time, iv) a well-grounded software data model.

Domain entities and design patterns

One of the main challenges of the DH community is to provide suitable software models and tools (Ciotti, 2014). To model the literary domain and the relative user requirements, we chose to follow the engineering principles of **object-oriented analysis and design** (Ackerman and Gonzalez, 2011). The digital representation of a textual resource is a challenge as it involves several theoretical and epistemological issues in semiotics, paleography, philology, linguistics, engineering, and computer science (McCarty, 2005; Meister, 2012; Moretti, 2013; Robinson, 2013; Sahle, 2013).

In this work, we define each textual element by means of four properties: i) the **version** allows to select a specific textual element among those available in its history of changes; ii) the **granularity** represents a level of a hierarchical structure (e.g. a page composed of lines); iii) the **position** provides the location of a textual element within the hierarchical representation (e.g. the second page of a book); iv) the **layer** indicates the set of homogeneous information the textual element belongs to (e.g. morphological layer). As pointed outby (Buzzetti, 2002; McGann, 2004; Pierazzo, 2015), the information conveyed by a textual resource is logically organized through multiple layers (also called *dimensions*) of information.

On these four properties we have designed and implemented a set of core entities as the fundamental data types shared among all the components of the environment (Fig. 1)[1]. The **Source** class is in charge of managing the low-level data: it is composed of a **Payload** representing the information conveyed by the textual resource and a **SourceType** which indicates the nature of the Source (e.g. text, image, audio, etc.). Payload objects (as used in networking) have the only purpose of encapsulating the information. The **Locus** and the **PlaceOfInterest** (POI) classes identify, through a *composition pattern*, specific data fragments of the source content, and they are used to establish the boundaries of an **Annotation**. A chunk of text, for example, can be addressed to by a locus having a POI (of type Sequence of Interest) representing its start and end coordinates. Similarly, a region of an image can be identified by a locus having a POI (of type Region of Interest) composed of a sequence of coordinates. The Locus and POI provide a stand-off text annotation technique able to tackle, for example, the overlapping hierarchies problem, which cannot be handled easily with inline markup techniques (Schmidt, 2010). As a matter of fact, it is possible, simultaneously, to manipulate a resource on the basis of its documental and textual structure (Renear et al., 1996; Robinson, 2013) (see the example in the following section). However, since stand-off models are affected by the issue of the indexing updating, a dedicated component must be in charge of automatically maintaining the coherence of the annotations each time the underlying text is edited.

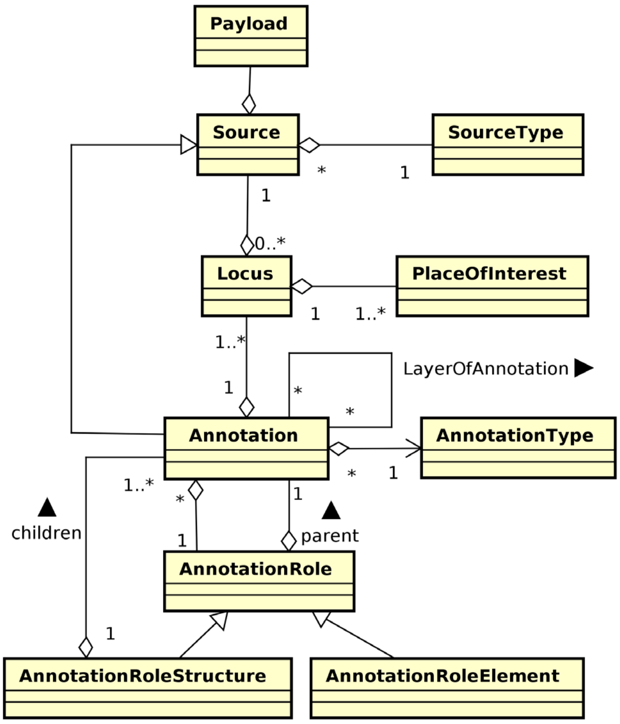
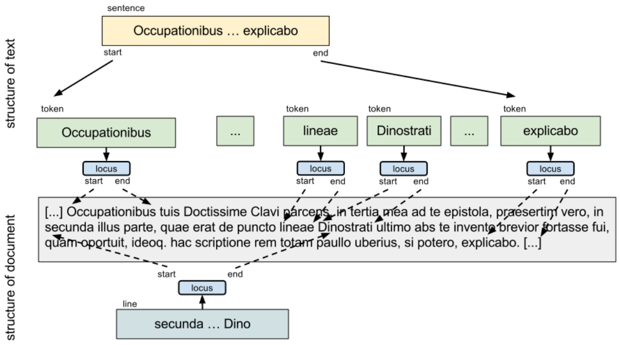


Fig. 1: Class diagram of the domain entities

An Annotation represents an information associated to a locus and is defined by an **AnnotationType** (e.g. a token, a lemma, a named entity, etc.). Since the hierarchical structure of the source may evolve over time, the changes to the relative tree must be managed. For example, a tree structure having tokens as leaves could need to be updated with a finer-grained layer of characters (e.g. to assign annotations to specific letters). In this case, the tokens should become intermediate nodes and the characters would become the leaf nodes. Typically, this kind of editing is unpredictable and it often implies heavy adaptations if the software is not flexible enough to manage changes in the underlying text representation schema. Consequently, we decided to exploit the flexibility of the Object Oriented model by adopting the Role Design Pattern (Fowler, 1997) to switch between leaf and intermediate nodes dynamically. This pattern has been implemented by the **AnnotationRole**, **AnnotationRoleElement** and **AnnotationRoleStructure** classes. Moreover, an annotation is a source in itself (see the inheritance relationship between the Annotation and the Source classes in Fig. 1) and, thus, it can be annotated recursively.

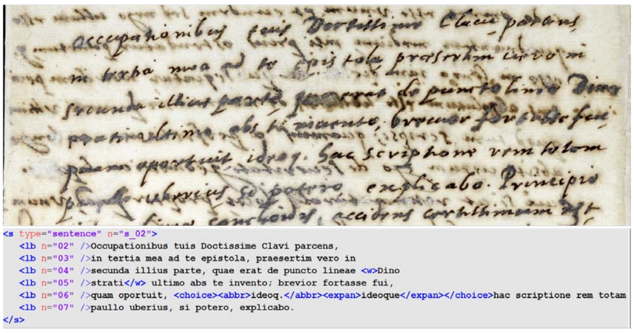
An Example

We here introduce an example showing a representation of a snippet of text with annotations. The chosen text is an excerpt of a letter, written in Latin, belonging to the epistolary corpus of the Clavius on the Web project[2] (Abrate et al., 2014b). Fig. 2 shows a typical way of encoding sentences and lines with a markup language as TEI (Burnard, 2014): the resulting XML hierarchical structure has been broken by the addition of the line anchors (<lb />) mixing up the textual and documental structure of the text. Indeed, to preserve the integrity of the word "Dinostrati" (spanning across lines 4 and 5), it is necessary to encapsulate it with the element <w />.



[...] Occupationibus tuis Doctissime Clavi parcens, in tertia mea ad te epistola, praesertim vero, in secunda illus parte, quae erat de puncto lineae Dinostrati ultimo abs te invento brevior fortasse fui, quam oportuit, ideoq. hac scriptione rem totam paullo uberius, si potero, explicabo. [...]

Fig. 2: A standard way of encoding a text with TEI-XML



```
<s type="sentence" n="s_02">
   <lb n="02" />Occupationibus tuis Doctissime Clavi parcens,
   <lb n="03" />in tertia mea ad te epistola, praesertim vero in
   <lb n="04" />secunda illius parte, quae erat de puncto lineae <w>Dino
   <lb n="05" />strati</w> ultimo abs te invento; brevior fortasse fui,
   <lb n="06" />quam oportuit, <choice><abbr>ideoq.</abbr><expan>ideoque</expan></choice>hac scriptione rem totam
   <lb n="07" />paullo uberius, si potero, explicabo.
</s>
```

Fig. 3: Multi-layered stand-off annotation of text

The model we propose solves this problem with the stand-off annotations: as shown in Fig. 3 the document (made of lines) and the textual structure (made of sentences and words) are logically separated. Lines, sentences and words do not overlap and they are structured in separate hierarchies.

Next Steps

We plan, in future works, to release a first version of a web environment, called *Omega*, built around the core entities that we here described. The environment will allow to load, index, annotate, and query a textual collection. Furthermore, we'll carry on the development of modules for text analysis and textual scholarship with the related APIs.

Bibliography

Abrate, M., Del Grosso, A. M., Giovannetti, E., Lo Duca, A., Luzzi, D., Mancini, L., Marchetti, A., Pedretti, I. and Piccini, S. (2014a). Sharing Cultural Heritage: the Clavius on the Web Project. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik*. European Language Resources Association (ELRA), pp. 627–34.

Abrate, M., Del Grosso,A. M., Giovannetti, E., Lo Duca, A., Marchetti, A., Mancini, L., Pedretti, I. and Piccini, S. (2014b). Il Progetto Clavius on the Web: tecnologie linguistico-semantiche al servizio del patrimonio documentale e degli archivi storici. In Rossi, F. and Tomasi, F. (eds), *Book of Abstracts of 30 AIUCD Conference, Bologna*.

Ackerman, L. and Gonzalez, C. (2011). *Patterns-Based Engineering: Successfully Delivering Solutions Via Patterns*. Addison-Wesley.

Albanesi, D., Bellandi, A., Benotto, G., Di Segni, G. and Giovannetti, E. (2015). When Translation Requires Interpretation: Collaborative Computer–Assisted Translation of Ancient Texts. *LaTeCH 2015*: 84–88.

Almas, B. and Beaulieu, M.-C. (2013). Developing a New Integrated Editing Platform for Source Documents in Classics. *Literary and Linguistic Computing*, **28**(4): 493–503 doi:10.1093/llc/fqt046.

Ashmore, S. and Runyan, K. (2014). *Introduction to Agile Methods*. Upper Saddle River, NJ: Addison-Wesley Professional, Pearson Education.

Bellandi, A., Albanesi, D., Bellusci, A., Bozzi, A. and Giovannetti, E. (2014). The Talmud System: a Collaborative web Application for the Translation of the Babylonian Talmud Into Italian. *The First Italian Conference on Computational Linguistics CLiC-It 2014*, pp. 53–57.

Bordalejo, B. and Robinson, P. (2015). A new system for collaborative online creation of Scholarly Editions in digital form. *1st Dixit Convension on Technology, Software, Standards for the Digital Scholarly Edition Workshop*. The Hague.

Boschetti, F. and Del Grosso, A. M. (2015). TeiCoPhiLib: A Library of Components for the Domain of Collaborative Philology. *Journal of the Text Encoding Initiative*(8). doi:10.4000/jtei.1285. http://jtei.revues.org/1285 (accessed 3 March 2016).

Boschetti, F., Del Grosso, A. M., Khan, A. F., Lamé, M. and Nahli, O. (2014). A top-down approach to the design of components for the philological domain. *Book of Abstract of Digital Humanities Conference (DH), Lausanne, Switzerland*. Alliance of Digital Humanities Organisations, pp. 109–11.

Bozzi, A. (2013). G2A: A Web application to study, annotate and scholarly edit ancient texts and their aligned translations. (Ed.) ERC Ideas 249431 *Studia Graeco-Arabica*, **3**: 159–71.

Bozzi, A. (2015). Greek into Arabic, a research Infrascructure based on computational modules to annotate and query historical and philosophical digital texts. Part I: Methodological aspects. In Bozzi, A. (ed), *Digital Texts, Translations, Lexicons in a Multi-Modular Web Application: Methods and Samples*. Firenze: Leo S. Olschki editore, pp. 27–42.

Burnard, L. (2014). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.9.1. http://www.tei-c.org/Guidelines/P5/index.xml (accessed 3 March 2016).

Buschmann, F., Henney, K. and Schmidt, D. C. (2007). *Pattern-Oriented Software Architecture, On Patterns and Pattern Languages*. (Pattern-Oriented Software Architecture). Hoboken: John Wiley & Sons.

Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P. and Stal, M. (1996). Pattern-oriented Software Architecture - A System of Patterns. J. Wiley and Sons Ltd., pp. 171–92.

Buzzetti, D. (2002). Digital Representation and the Text Model. *New Literary History*, **33**(1): 61–88.

Causer, T. and Terras, M. (2014). "Many hands make light work. Many hands together make merry work": Transcribe Bentham and crowdsourcing manuscript collections.

Ciotti, F. (2014). Digital Literary and Cultural Studies: State of the Art and Perspectives. *Between*, **4**(8). doi:10.13125/2039-6597/1392. http://dx.doi.org/10.13125/2039-6597/1392 (accessed 3 March 2016).

Cohn, M. (2004). *User Stories Applied: For Agile Software Development*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc.

Collins-Cope, M., Rosenberg, D. and Stephens, M. (2005). *Agile Development with ICONIX Process: People, Process, and Pragmatism*. Berkely, CA, USA: Apress.

Fowler, M. (1997). Dealing with roles. *Proceedings of the International Conference on Pattern Languages of Programs*, vol. 97, pp. 13–37.

Gamma, E., Helm, R., Johnson, R. and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S. and Piazza, F. (2013). Pundit: augmenting web contents with semantics. *Literary and Linguistic Computing*, **28**(4): 640–59.

Grill, T., Polacek, O. and Tscheligi, M. (October 29-312012). Methods Towards API Usability: A Structural Analysis of Usability Problem Categories. *Proceedings of the 4th International Conference on Human-Centered Software Engineering, Toulouse, France*. Berlin, Heidelberg: Springer-Verlag, pp. 164–80. doi:10.1007/978-3-642-34347-6_10.

Del Grosso, A. M. (2013). Indexing techniques and variant readings management. (Ed.) D'Ancona, C. *Studia Graeco-Arabica*, **3**: 209–30.

Del Grosso, A. M. and Nahli, O. (2014). Towards a flexible open-source software library for multi-layered scholarly textual studies: An Arabic case study dealing with semi-automatic language processing. *Proceedings of 3rd IEEE International Colloquium, Information Science and Technology (CIST), Tetouan, Marocco*. Washington, DC, USA: IEEE, pp. 285–90. doi:10.1109/CIST.2014.7016633.

Hedges, M., Neuroth, H., Smith, K. M., Blanke, T., Romary, L., Küster, M. and Illingworth, M. (2013). TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship. *Journal of the Text Encoding Initiative*(5). doi:10.4000/jtei.774 (accessed 3 March 2016).

Hinrichs, E., Hinrichs, M. and Zastrow, T. (2010). WebLicht: Web-based LRT services for German. *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, pp. 25–29.

Houghton, H., Sievers, M. and Smith, Catherine (2014). The Workspace for Collaborative Editing. *Digital Humanities 2014*. Laussanne: Alliance of Digital Humanities Organisations, pp. 204–05.

D'Iorio, P. (2015). On the scholarly use of the Internet, a conceptual model. In Bozzi, A. (ed), *Digital Texts, Translations, Lexicons in a Multi-Modular Web Application: Methods and Samples*. Firenze: Leo S. Olschki editore, pp. 1–25.

Jordanous, A., Lawrence, K. F., Hedges, M. and Tupman, C. (June 13-152012). Exploring Manuscripts: Sharing Ancient Wisdoms Across the Semantic Web. *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS), Craiova, Romania*. New York, NY, USA: ACM, pp. 44:1–44:12. doi:10.1145/2254129.2254184.

McCarty, W. (2005). *Humanities Computing*. Palgrave Macmillan.

McCarty, W. (2008). Signs of times present and future. *Human Discussion Group*, **22**(218).

McGann, J. (2004). Marking Texts of Many Dimensions. In Schreibman, S., Siemens, R. and Unsworth, J. (eds), *A Companion to Digital Humanities*. (Blackwell Companions to Literature and Culture). Blackwell Publishing Ltd, pp. 198–217.

Meister, J. C. (2012). DH is us or on the unbearable lightness of a shared methodology. *Historical Social Research*, **37**(3): 77–85.

Moretti, F. (2013). *Distant Reading*. Verso Books.

Moretti, G., Tonelli, S., Menini, S. and Sprugnoli, R. (2014). ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment. *Proceedings of the First Italian Conference on Computational Linguistics (CLIC-2014)*. Pisa, Italy.

Moulin, C., Nyhan, J., Ciula, A., Kelleher, M., Mittler, E., Tadić, M., Ågren, M., Bozzi, A. and Kuutma, K. (2011). *Research Infrastructures in the Digital Humanities*. http://www.esf.org/hosting-experts/scientific-review-groups/humanities-hum/strategic-activities/research-infrastructures-in-the-humanities.html.

Nowviskie, B. (2007). Collex: Facets, Folksonomy, and Fashioning the Remixable web. *Book of Abstract of Digital Humanities Conference (DH), University of Illinois at Urbana-Champaign*. Alliance of Digital Humanities Organisations.

Ott, W. (2000). Strategies and tools for textual scholarship: the Tübingen system of text processing programs (TUSTEP). *Literary and Linguistic Computing*, **15**(1): 93–108. doi:10.1093/llc/15.1.93. http://llc.oxfordjournals.org/content/15/1/93.abstract.

Ott, W. and Ott, T. (2014). Critical Editing with TXSTEP. In Terras, M. (ed), *Book of Abstracts of the Digital Humanities Conference, Lausanne, Switzerland*. Alliance of Digital Humanities Organisations, pp. 509–13.

Paradis, J., Fendt, K., Kelley, W., Folsom, J., Pankow, J., Graham, E. and Subbaraj, L. (2013). Annotation Studio: Bringing a Time-Honored Learning Practice into the Digital Age. *Whitepaper*. http://cmsw.mit.edu/annotation-studio-whitepaper/ (accessed 3 March 2016).

Pierazzo, E. (2015). *Digital Scholarly Editing : Theories, Models and Methods*. Farnham Surrey: Ashgate.

Renear, A. H., Mylonas, E. and Durand, D. (1996). Refining our notion of what text really is: The problem of overlapping hierarchies. (Ed.) Hockey, S. M. *Research in Humanities Computing*, **4**: 263–80.

Robinson, P. (2013). Towards a theory of digital editions. (Ed.) Mierlo, W. V. and Fachard, A. *Variants*, (10): 105–31.

Ruimy, N., Piccini, S. and Giovannetti, E. (2012). Defining and Structuring Saussure's Terminology. In Fjeld, R. V. and Torjusen, J. M. (eds), *Proceedings of 15th EURALEX International Congress*. Oslo, Norway, Department of Linguistics and Scandinavian Studies, University of Oslo, Reprosentralen: UiO press, pp. 828–33.

Sahle, P. (2013). *Digitale Editionsformen: Teil 3: Textbegriffe Und Recodierung; Zum Umgang Mit Der Überlieferung Unter Den Bedingungen Des Medienwandels*. Vol. 3. BoD–Books on Demand.

Schmidt, D. (2010). The inadequacy of embedded markup for cultural heritage texts. *Literary and Linguistic Computing*, **25**(3): 337–56. doi:10.1093/llc/fqq007.

Schmidt, D. (2014). Towards an Interoperable Digital Scholarly Edition. *Journal of the Text Encoding Initiative*(7). doi:10.4000/jtei.979.

Sinclair, S. and Rockwell, G. (2012). the Voyant Tools Team (web application) *Voyant Tools*. http://voyant-tools.org (accessed 3 March 2016).

Steiner, C., Agosti, M., Sweetnam, M., Hillemann, E.-C., Orio, N., Ponchia, C., Hampson, C., et al. (2014). Evaluating a digital humanities research environment: the CULTURA approach. *International Journal on Digital Libraries*, **15**(1): 53–70. doi:10.1007/s00799-014-0127-x.

Terras, M. and Crane, G. (eds). (2010). *Changing the Center of Gravity: Transforming Classical Studies through Cyberinfrastructure*. Piscataway: Gorgias Press.

Tulach, J. (2008). *Practical API Design: Confessions of a Java Framework Architect*. 1st ed. Berkely, CA, USA: Apress.

Notes

[1] The ongoing implementation of the environment is available at: https://github.com/literarycomputinglab

[2] Clavius on the Web is a project funded by Registro.it and partecipated by the Institute of Informatics and Telematics (IIT-CNR), the Institute of Computational Linguistics "A. Zampolli" (ILC-CNR), and the Historical Archives of the Pontifical Gregorian University (APUG). Website: http://claviusontheweb.it/

EVI-LINHD. A Virtual Research Environment for the Spanish-speaking Community

Gimena del Rio Riande
gdelrio.riande@gmail.com
CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), LINHD-UNED (Laboratorio de Innovación en Humanidades Digitales, UNED)

Elena González-Blanco García
egonzalezblanco@flog.uned.es
LINHD-UNED (Laboratorio de Innovación en Humanidades Digitales, UNED)

Clara Martínez Cantón
cimartinez@flog.uned.es
LINHD-UNED (Laboratorio de Innovación en Humanidades Digitales, UNED)

Juan José Escribano
juanjo.escribano@gmail.com
LINHD-UNED (Laboratorio de Innovación en Humanidades Digitales, UNED)

Although Digital Humanities have been defined from a discipline perspective in many ways, it is surely a field still looking for its own objects, practices and methodologies. Their development in the Spanish-speaking countries is no exception to this process and, even it is complex to trace a unique genealogy to give account for the evolving field in Spain and Latin America (Gonzalez-Blanco, 2013; Spence and Gonzalez-Blanco, 2014; Rio Riande 2014a, 2014b), the emergence of various associations in Mexico (RedDH), Spain (HDH) and Argentina (AAHD) that seek for a constant dialogue (Galina, González-Blanco and Rio Riande, 2015), and academic lab and DH center initiatives such as LINHD (Spain and Argentina), GRINUGR (Spain), Medialab USAL, LABTEC (Argentina), TadeoLab (Colombia), Elabora HD (Mexico), among others, make it clear that research has become increasingly "global, multipolar and networked" (Llewellyn Smith, et al., 2011) and that the academic field is looking for a global outreach and aims to open spaces of shared virtual work. Virtual Research Communities (VRCs) are a consequence of these changes.

Virtual Research Environments (VREs) have become central objects for digital humanist community, as they help global, interdisciplinary and networked research taking of profit of the changes in "data production, curation and (re-)use, by new scientific methods, by changes in technology supply" (Voss and Procter, 2009: 174-90). DH Centers, labs or less formal structures such as associations benefit from many kind of VREs, as they facilitate researchers and users a place to develop, store, share and preserve their work, making it more visible. The focus and implementation of each of these VREs is different, as Carusi and Reimer (2010) show in their comparative analysis, but there are some common guidelines, philosophy and standards that are generally shared (as an example, see the Centernet map and guidelines of TGIR Huma-Num, 2015).

This poster presents the structure and design of the VRE of LINHD, the Digital Innovation Lab at UNED (http://linhd.uned.es), and the first Digital Humanities Center in Spain. This VRE focuses on the possibilities of a collaborative environment for (profane or advanced) Spanish-speakers scholarly digital editors. Taking into account the language barrier that English may suppose for a Spanish-speakers scholar or student and the distance they may encounter with the data and organization of the interface (in terms of computational knowledge) while facing a scholarly digital edition or collection, LINHD's VRE comes as a solution for the VRC interested in scholarly digital work. Moreover, it will make it possible to add an apply tools that contribute to improve Spanish-English applications or tools developed locally, such as *Contawords*, by Iula-UPF http://contawords.iula.upf.edu/executions. Opening such an environment to the Spanish speaking world will make it possible to reach different kinds of communities, whose profile and training in digital humanities differ from the typical users of DH tools and environment. Testing all these tools in this new environment will, for sure, draw interesting project results.

In this sense, our project dialogues and aims to join the landscape of other VREs devoted to digital edition, such as *Textgrid*, *e-laborate*, etc. and, in a further stage, to build a complete virtual environment to collect and classify data, tools and projects, work and publish them and share the results with the research community. After having studied the structure and components of other digital virtual environment, our VRE has been designed on a humanist-user centered perspective, in which interface design, accessibility easiness and familiarity with tools and standards are key factors.

Therefore, the key of our VRE is the combination of different open-source software that will enable users to complete the whole process of developing a digital editorial project. The environment is, up-to-now, divided into three parts: 1) A repository of data to (projects, tools, etc.) with permanent identifiers in which the information will be indexed through a semantic structured ontology of metadata and controlled vocabularies (such as *Isidore* and *Huni*, but using *LINDAT* software by *Clarin. eu*). 2) A working space based on the possibilities of *eXistDB* to work on text encoding together with *Tei-Scribe*, a tool developed at LINHD to tag texts in an intuitive way, storing and querying, plus some publishing tools (pre-defined stylesheets and some other open-source projects, such as *Sade*, *Versioning machine*, etc.). 3) A collaborative cloud workspace which integrates a wiki, a file archiving system and a publishing space for each team.

Sustainability and long-term preservation are issues which we contemplate from the beginning, as our group is leading the addition of Spain into Dariah and LINHD is also part of a Clarin-Knowledge center with two powerful NLP groups from U.Pompeu Fabra in Barcelona and IXA in País Vasco. Our project has been conceived according to DH standards and open-source tools and its infrastructure is supported by our university UNED.

Bibliography

Candela, L. Virtual Research Environments. GRDI2020. http://www.grdi2020.eu/Repository/FileScaricati/eboe-8fea-c496-45b7-a0c5-831b90fe0045.pdf (accessed 28-10-2015).

Carusi, A. and T. Reimer, (2010). Virtual Research Environment Collaborative Landscape Study. *A JISC funded project*. Oxford e-Research Centre, University of Oxford and Centre for e-Research, King's College London https://www.jisc.ac.uk/rd/projects/virtual-research-environments (accessed 28-10-2015).

Galina, I., González Blanco García, E. and Rio Riande, G. del (2015). Se habla español. Formando comunidades digitales en el mundo de habla hispana. *Abstracts of the HDH 2015 Conference*, Madrid, Spain. http://hdh2015.linhd.es/ebook/hdh15-galina.xhtml (accessed 28-10-2015).

González-Blanco Garcí A., E. (2013). Actualidad de las Humanidades Digitales y un ejemplo de ensamblaje poético en la red: ReMetCa. *Cuadernos Hispanoamericanos*, **761**: 53-67.

Llewellyn Smith, C., Borysiewicz, L., Casselton, L., Conway, G., Hassan, M., Leach, M., et al. (2011). *Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century*. London: The Royal Society.

Rio Riande, G. del (2014a). ¿De qué hablamos cuando hablamos de Humanidades Digitales? *Abstracts of the AAHD Conference. "Culturas, Tecnologías, Saberes* Buenos Aires, Argentina. http://www.aacademica.com/jornadasaahd/toc/6?abstracts (accessed 28-10-2015).

Rio Riande, G. del (2014b). ¿De qué hablamos cuando hablamos de Humanidades Digitales? http://blogs.unlp.edu.ar/didacticaytic/2015/05/04/de-que-hablamos-cuando-hablamos-de-humanidades-digitales/. (accessed 28-10-2015).

Spence, P. and González-Blanco, E. (2014). A historical perspective on the digital humanities in Spain, *H-Soz-Kult*, doi: 22.10.2014,http://www.hsozkult.de/text/id/texte-2535.

The Status Quo of Digital Humanities in Europe, *H-Soz-Kult*, doi: 22.10.2014. (accessed 28-10-2015).

Tgir H.-N. (2011). *Le guide des bonnes pratiques numériques*. http://www.huma-num.fr/ressources/guide-des-bonnes-pratiques-numeriques (version of 13-1-2015). (accessed 28-10-2015).

Voss, A. and Procter, R. (2009). Virtual research environments in scholarly work and communications, *Library Hi Tech*, **27**(2): 174–90.

Diachronic Semantic Lexicon of Dutch (Diachroon semantisch lexicon van de Nederlandse Taal; DiaMaNT)

Katrien A. C. Depuydt
katrien.depuydt@inl.nl
Instituut voor Nederlandse Lexicologie, Netherlands, The

Dutch language has been described extensively in the comprehensive historical dictionaries of the Institute for Dutch lexicology. These dictionaries (Oudernederlands Woordenboek, Dictionary of Old Dutch, ca. 500-1200; Vroegmiddelnederlands Woordenboek, Dictionary of Early Middle Dutch, 1200-1300 ; Middelnederlandsch Woordenboek; MNW, Dictionary of Middle Dutch, ~1250-550; Woordenboek der Nederlandsche Dictionary of the Dutch Language, 1500-976) cover over 15 centuries of Dutch and are as such a perfect guide to understanding historical language. The dictionaries also provide the core material for the diachronic computational lexicon of Dutch (GiGaNT), that can be used to support search in historical texts by users without (expert) knowledge of historical spelling variation: when searching for *slager* ('butcher') the user also gets the morphological and spelling variants like *slagers, slagher(s), slaeger(s) slaegher(s)* or *slegher(s)*. However, when a user wants to study the history of the butcher's trade, it is not immediately obvious from the way these traditional dictionaries are structured that one has also to look for *vleeschhouwer* or *beenhouwer* or *beenhakker.* And it is only after reading the complete articles that a user learns that *vleeschouwer* can also mean 'executioner', and *slager* 'a person who slays so.', be it though that in the case of *vleeschhouwer* the meaning *'executioner´* is derived from vleeschhouwer 'butcher', while *slager* in contemporary meaning 'butcher' is derived from the meaning *'a person who slays so'.*

In this contribution we describe the first results of our work on the development of a diachronic semantic lexicon of Dutch. The lexicon aims to enhance text accessibility and to foster research in the development of concepts, by interrelating attested word forms and semantic units (concepts), and tracing semantic developments in time. In the lexicon, the diachronic onomasiology, i.e. the change in naming of concepts and the diachronic semasiology, i.e. the change in meaning of words, will be recorded in a way suitable for use by humans and computers. The onomasiological part of the lexicon is meant to enhance recall in text retrieval by providing different verbal expressions of a concept or related concepts (*slager → beenhouwer, beenhakker, vleeshouwer; boer → landman*). The diachronic semasiological component (which charts semantic change), aims to enhance precision by enabling the user to take semantic change into account; the oldest meaning of *appel* for example is 'a fruit' (so *appel* is also used for pears, plums etc.).

We describe the structure of the diachronic semantic lexicon and procedures for the acquisition and aggregation of content. The INL historical dictionaries will be the main source of the lexicon, as these dictionaries describe the Dutch lexicon from the 6th to the 20th century and cover most of the basic vocabulary of this period. Word sense descriptions are illustrated by dated quotations, which constitute a first step towards dating a concept. The temporal distribution of quotations pertaining to different senses gives a first picture of the diachronic development of the sense inventory of a headword. The fact that many words in the historical dictionaries are defined (partly) by synonym definitions and contemporary semantic (near)-equivalents enables us to extract an initial set of semantic relations.

Information from other sources is not disregarded. For contemporary Dutch, several lexical resources cataloguing semantic relationships are available. This includes traditional synonym dictionaries like Brouwers "Het Juiste woord" and more recent initiatives such as Open Dutch Wordnet (Vossen). For some specific domains, thesauri with a diachronic component are in development (eg. the *HISCO* (http://historyofwork.iisg.nl/index.php)).

Besides lexical sources, diachronic corpus material[1] and corpus-based methods are no less essential to the development and verification of the relevance of the lexicon content. This includes: i) corpus based analysis of semantic change at the "type"-level, using distributional methods. Here, the fact that our starting point is defined by the set of quotation dates per word sense provides an interesting perspective. ii) research into the application of token-based distributional methods to the interlinking of historical corpora and lexical resources.

Bibliography

Fellbaum, C. ed. (1999). *WORDNET. An Electronic Lexical Database.* London: The MIT Press.

Geeraerts, D., et al. (1994). *The Structure of Lexical Variation. Meaning, Naming, and Context.* Berlin/New York: Mouton de Gruyter.

Geeraerts, D. (1997). *Diachronic Prototype Semantics. A Contribution to Historical Lexicology.* Oxford: Clarendon Press.

Geeraerts, D. (2010). *Theories of Lexical Semantics.* Oxford/New York: Oxford University Press.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop*, pp. 67-71.

Heylen, K., et al. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, **157**: 153-72.

Kay, C. J. and Chase, T. J. P. (1987). Constructing a Thesaurus database. *Literary and Linguistic computing*, **2**(3): 161-63.

Laurence, S. and Margolis, E. (1999). Concepts and Cognitive Science. In Margolis, E. and Laurence, S., *Concepts. Core Readings.* Cambridge (US)/London: The MIT Press, pp. 3-81.

Sijs, N. van der (2001). *Etymologie in het digitale tijdperk. Een chronologisch woordenboek als praktijkvoorbeeld.* Ph.D. thesis, Universiteit Leiden.

Vanhove, M. ed. (2008). *From Polysemy to Semantic Change. Towards a typology of lexical semantic associations.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Vossen, P. ed. (1998). *EuroWordNet: A mulitlingual database with lexical semantic networks.* Reprinted from *Computer and the Humanities*, **Vol.** 32, Nos. 2-3, 1998. Dordrecht/Boston/London: Kluwer Academic Publishers.

Notes

[1] Corpora: DBNL (digital library of Dutch literature, http://www.dbnl.nl), digitized newspaper collections at the Dutch Royal Library, and other collections digitized by the Royal Library (http://www.delpher.nl).

A Web-based Tool Called Gauntlet: From Iterative Design To Interactive Drawings Annotation

Gregory Dessart
gregory.dessart@unil.ch
University of Lausanne, Faculty of Theology and Religious Studies: Institute for Social Sciences of Contemporary Religions, Switzerland

Martial Sankar
Martial.Sankar@isb-sib.ch
Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland

Anastasia Chasapi
Anastasia.Chasapi@unil.ch
Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland

Guido Bologna
Guido.Bologna@unil.ch
Computer Vision and Multimedia Laboratory, University of Geneva, Switzerland

Zhargalma Dandarova Robert
Zhargalma.Dandarova@unil.ch
University of Lausanne, Faculty of Theology and Religious Studies: Institute for Social Sciences of Contemporary Religions, Switzerland

Pierre-Yves Brandt
Pierre-Yves.Brandt@unil.ch
University of Lausanne, Faculty of Theology and Religious Studies: Institute for Social Sciences of Contemporary Religions, Switzerland

Gauntlet was developed for the international SNSF project "Drawings of gods", which is dealing with up to 5'100 drawings collected in different parts of the world (e.g. Japan, Russia, Switzerland, Romania). The data was digitalized and deposited on a publicly available database designed on MySQL for this project (http://ddd.unil.ch). The presentation of this tool will then make use of a specific application on such material.

Subsequent analyses involve interpretation of collected data and metadata (Brandt et al., 2009; Dandarova, 2013) as well as image analysis (Konyushkova et al., submitted). While algorithms and software tools for automated image analysis of photographic pictures are steadily expanding, high-variability and approximate nature of children's drawings make analysis and object detection (e.g. stars, clouds, eyes) rather challenging.

Convinced that the curation (i.e. manual annotation) of our complex objects is the solution both to tackling such a big data problematic and to guiding further automated computational approaches, we built Gauntlet, a web application tool for image annotation. It results from a collaborative and interdisciplinary software development process gathering psychologists, theologians and bioinformatics specialists. The tool is available for Chrome (version 44.0 or higher) and Firefox (version 41.0.2 or higher). It was built on modern and open source web standards for software development such as HTML5/CSS3 and AngularJS. There is no restriction to the use of this tool and it is GNU GPLv>=2. The logic behind follows from web tool development in bioinformatics, aiming to provide platform-independent and barrier-free solutions - which could not be found with existing tools.

Gauntlet's main benefit is to provide our worldwide collaborators with common semantics from a set of attributes (i.e. features) and geometrical tools for annotating the collected drawings. The set is displayed on the interface as a hierarchical list - termed "annotation tree" - of categories and subcategories (e.g. "characters area" branch contains sub-branch attributes: "human", "non-human", "text" and "blank sheet"). The tree is designed and edited via Excel and converted to JSON for web display. Such an implementation grants us flexibility and editing of the annotation tree based on curators' comments (full text comments can be dropped at every level) and the degree of attributes occurrences. The annotation tree first took after McCarty et al.'s personification model (2004) approach, but is currently fashioned in such a way that the role played by annotators' subjectivity should be reduced.

While it allows relatively fast annotations of large samples of drawings, Gauntlet's main purpose is to provide positional coordinates of attributes from smooth and accurate user-image interactions. Such interactions rely on points to target repeated objects of a specific attribute (e.g. accessories or hands) and boxes to mark off the whole spanning area of an attribute (e.g. characters area). Curators can, at any time, export the current status of annotation in CSV format and run comparative statistical analysis on feature frequency, occupancy or positioning. So far, annotations have been carried out by researchers who are affiliated with our project, although crowdsourcing may be used in the future.

In the end, it is expected to help highlight developmental, intercultural and interfaith variations on the material collected for the abovementioned project "Drawings of gods". Moreover, exports from Gauntlet will help complete further explorations into pattern recognition by providing the ground truth.

At this stage the tool is client-based and main features are ready for use. Even though it is stand-alone, it may go through further development to plug into API for operating with any kind of data storage systems (e.g. RDBMS, nosql or rdf-based). Beside this, due to its flexibility such a tool might prove practical for a wider use, targeting the processing of data in various contexts, such as photos of paintings or other art forms.

Overall, the collaborative and interdisciplinary work from which Gauntlet emerged represents a great deal of mutually enriching influences between agents from various fields ready to revise their initial views, thus shedding light on what today's scientific project development should look like.

Bibliography

Brandt, P.-Y., Kagata Spitteler, Y. and Gillièron Paléologue, C. (2009). La représentation de Dieu: Comment les enfants japonais dessinent Dieu. *Archive de Psychologie*, **74**: 171-203.

Dandarova, Z. (2013). Le dieu des enfants: Entre l'universel et le contextuel. In Brandt, P.-Y. and Day, J. M. (eds.), *Psychologie du développement religieux: Questions classiques et perspectives contemporaines*, Labor et Fides, pp. 159-87.

Konyushkova, K., Arvanitopoulos, N., Süsstrunk, S., Dandarova, Z., and Brandt, P.-Y. (submitted). God(s) know(s): Developmental and cross-cultural patterns in children drawings. *Journal on Computing and Cultural Heritage*.

McCarty, W., Matthews, M., Suksi, A., Wright, B. and Bradley, J. (2004). An Analytical Onomasticon to the Metamorphoses of Ovid. *Classical Studies Publications*.

EVT 2.0: a new architecture for critical editions in digital form

Chiara Di Pietro
dipi.chiara@gmail.com
University of Pisa, Italy

Chiara Alzetta
chiara.alzetta@gmail.com
University of Pisa, Italy

Ilaria Tiezzi
ilaria.tiezzi@gmail.com
University of Pisa, Italy

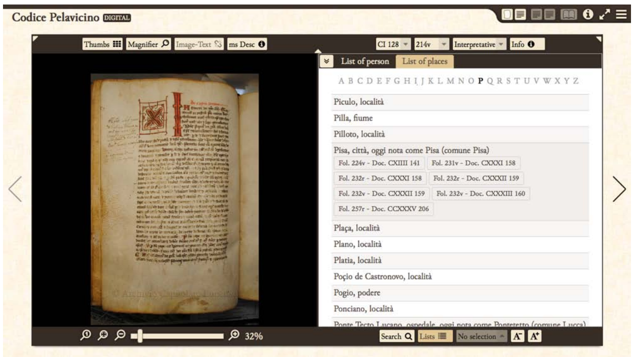
Raffaele Masotti
raffaele.masotti@gmail.com
University of Pisa, Italy

Roberto Rosselli Del Turco
roberto.rossellidelturco@unito.it
University of Turin, Italy

EVT (Edition Visualization Technology[1]) is a lightweight, open source tool specifically designed to create digital editions from XML-encoded texts, freeing the scholar from the burden of web programming and enabling the final user to browse, explore and study digital editions by means of a user-friendly interface, providing a set of tools (zoom, magnifier and hot-spots for manuscript images, text-image linking and an internal search engine for the edited texts) for research purposes. The starting point of the system is one or more documents in the standard TEI P5 format, which is turned into a web based application – a mix of HTML5, CSS3 and JavaScript – that can be easily shared on the Web. The text can be presented in different levels of edition (e.g. interpretative, diplomatic) and, besides the default visualization layout where text and scans of the original manuscript are linked together and placed side by side, a book reader mode can be activated if double side images are supplied.

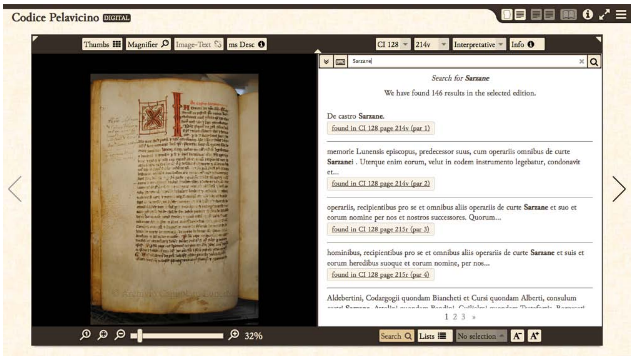
Even if it was born in the context of the specific use case the Digital Vercelli Book project, whose first version has been available online for about a year[2], it is able to fit different texts and needs. For example, it is now being used to publish the digital edition of the Codice Pelavicino manuscript[3], a medieval codex preserving charters dating back to the XIII century. Moreover, it has been used by the CVCE[4] (Centre Virtuel de la Connaissance sur l'Europe) as a starting point to create a publication framework for bilingual (French, English) documents of the W.E.U. (Western European Union). As a first consequence of these collaborations, EVT has been enriched with several new features[5]:

- thanks to an appropriate encoding, it is now possible to automatically create a list of all the entity names in the text that can be used both to provide a direct access to the page where a particular entity occurs and to present all the additional information considered relevant for that specific entity;
- an internal search engine allows the user to perform textual searches in a specific edition level for the shown text, and a virtual keyboard holding special characters is available if necessary;
- a generalized method to add critical textual notes has been implemented;
- all the information included in the TEI header, in particular the <msDesc> element, and in the <front> element, can now be shown to the user in a separate formatted panel;
- the user interface is now more flexible thanks to new configuration options, and it can be localized in different languages.



Piculo, località
Pilla, fiume
Pilloto, località
Pisa, città, oggi nota come Pisa (comune Pisa)
Fol. 224v - Doc. CXIIII 141 | Fol. 231v - Doc. CXXXI 158
Fol. 232r - Doc. CXXXI 158 | Fol. 232r - Doc. CXXXII 159
Fol. 232v - Doc. CXXXII 159 | Fol. 232v - Doc. CXXXIII 160
Fol. 257r - Doc. CCXXXV 206
Plaça, località
Plano, località
Platia, località
Poçio de Castronovo, località
Pogio, podere
Ponciano, località
Ponte Tecto Lucano, ospedale, oggi nota come Pontetetto (comune Lucca)

Figure 1: EVT 1.0 - List of named entities



Sarzane

Search for ***Sarzane***

We have found 146 results in the selected edition.

De castro **Sarzane**.
found in CI 128 page 214v (par 1)

memorie Lunensis episcopus, predecessor suus, cum operariis omnibus de curte **Sarzanei** . Uterque enim eorum, velut in eodem instrumento legebatur, condonavit et...
found in CI 128 page 214v (par 2)

operariis, recipientibus pro se et omnibus aliis operariis de curte **Sarzane** et suo et eorum nomine per nos et nostros successores. Quorum...
found in CI 128 page 215r (par 3)

hominibus, recipientibus pro se et omnibus aliis operariis de curte **Sarzane** et suis et eorum heredibus suoque et eorum nomine, per nos...
found in CI 128 page 215r (par 4)

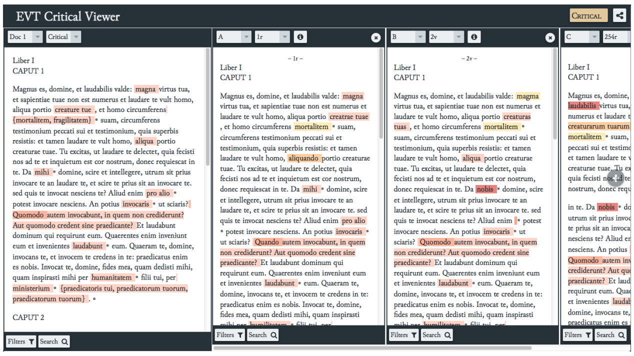
Aldebertini, Codargogii quondam Biancheti et Cursi quondam Alberti, consulum

Figure 2: EVT 1.0 - Search results

The continuous development and need to adapt EVT to different types of documents and TEI-encoded texts has shifted the development focus towards creation of a more general tool for the web publication of TEI-based documents, able to cater for multiple use cases. This decision has led to a complete revision of the infrastructure in order to improve flexibility and modularity, to make it easier to implement new features and to adjust the UI layout for different kinds of editions. This is why the development team decided to refactor the whole code of the viewer and base it on AngularJS[6], a Javascript framework which implements the MVC (Model View Controller) pattern to separate presentation, data, and logic component, providing a great modularity of the web application.

Perhaps the most important new feature developed for the next EVT version is the support for critical editions, again encoded according to the TEI XML P5 standard. This new level of edition is based on the current TEI relevant CA module and Guidelines chapter[7] and it supports the Parallel Segmentation method. The current implementation, however, is kept as generic and flexible as possible to make it easier to update it when the TEI module will be rewritten and expanded to become more powerful and suitable to philologists[8]. Among the different tools offered, EVT will provide a straight and quick link from the critical apparatus to the textual context of a specific reading; moreover, it will allow to compare witnesses' texts among each other or with respect to the base text of the edition (or to another specific text); finally, if the digitized images of each manuscript are provided, each variant can be examined in its palaeographic context.

From the point of view of the editor, the new architecture will be as easy to use as the current one: the only technical skill required from the editors will be a general competence in XML editing to configure EVT properly and the care to place each XML-related component of the edition (mainly the schema besides the encoded texts) into the correct area of the directory structure. Moreover, the editor will be able to modify the CSS style sheets to customize all aspects of text visualization, according to his/her needs.



Liber I
CAPUT 1

Magnus es, domine, et laudabilis valde: magna virtus tua, et sapientiae tuae non est numerus et laudare te vult homo, aliqua portio creature tue , et homo circumferens {mortalitem, fragilitatem} • suam, circumferens testimonium peccati sui et testimonium, quia superbis resistis: et tamen laudare te vult homo, aliqua portio creaturae tuae. Tu excitas, ut laudare te delectet, quia fecisti nos ad te et inquietum est cor nostrum, donec requiescat in te. Da mihi • domine, scire et intellegere, utrum sit prius invocare te an laudare te, et scire te prius sit an invocare te. sed quis te invocat nesciens te? Aliud enim pro alio • potest invocare nesciens. An potius invocaris • ut sciaris? Quomodo autem invocabunt, in quem non crediderunt? Aut quomodo credent sine praedicante? Et laudabunt dominum qui requirunt eum. Quaerentes enim inveniunt eum et invenientes laudabunt • eum. Quaeram te, domine, invocans te, et invocem te credens in te: praedicatus enim es nobis. Invocat te, domine, fides mea, quam dedisti mihi, quam inspirasti mihi per humanitatem • filii tui, per ministerium • {praedicatoris tui, praedicatorum tuorum, praedicatorum tuorum} . •

CAPUT 2

– 1r –
Liber I
CAPUT 1

Magnus es, domine, et laudabilis valde: magna virtus tua, et sapientiae tuae non est numerus et laudare te vult homo, aliqua portio creatrae tuae , et homo circumferens mortalitem • suam, circumferens testimonium peccati sui et testimonium, quia superbis resistis: et tamen laudare te vult homo, aliquando portio creaturae tuae. Tu excitas, ut laudare te delectet, quia fecisti nos ad te et inquietum est cor nostrum, donec requiescat in te. Da mihi • domine, scire et intellegere, utrum sit prius invocare te an laudare te, et scire te prius sit an invocare te. sed quis te invocat nesciens te? Aliud enim pro alio • potest invocare nesciens. An potius invocaris • ut sciaris? Quando autem invocabunt, in quem non crediderunt? Aut quomodo credent sine praedicante? Et laudabunt dominum qui requirunt eum. Quaerentes enim inveniunt eum et invenientes laudabunt • eum. Quaeram te, domine, invocans te, et invocem te credens in te: praedicatus enim es nobis. Invocat te, domine, fides mea, quam dedisti mihi, quam inspirasti mihi per humilitatem • filii tui, per

– 2v –
Liber I
CAPUT 1

Magnus es, domine, et laudabilis valde: magma virtus tua, et sapientiae tuae non est numerus et laudare te vult homo, aliqua portio creaturas tuas , et homo circumferens mortalitem • suam, circumferens testimonium peccati sui et testimonium, quia superbis resistis: et tamen laudare te vult homo, aliqua portio creaturae tuae. Tu excitas, ut laudare te delectet, quia fecisti nos ad te et inquietum est cor nostrum, donec requiescat in te. Da nobis • domine, scire et intellegere, utrum sit prius invocare te an laudare te, et scire te prius sit an invocare te. sed quis te invocat nesciens te? Aliud enim • potest invocare nesciens. An potius invocaris • ut sciaris? Quomodo autem invocabunt, in quem non crediderunt? Aut quomodo credent sine praedicante? Et laudabunt dominum qui requirunt eum. Quaerentes enim inveniunt eum et invenientes laudabunt • eum. Quaeram te, domine, invocans te, et invocem te credens in te: praedicatus enim es nobis. Invocat te, domine, fides mea, quam dedisti mihi, quam inspirasti mihi per humilitatem • filii tui, per

Liber I
CAPUT 1

Magnus es, domine,
laudabilis virtus tua,
numerus et laudare t
creaturarum tuarum
mortalitem • suam,
peccati sui et testimo
et tamen laudare te v
creaturae tuae. Tu ex
quia fecisti nos ad te
nostrum, donec requ

in te. Da nobis • do
utrum sit prius invoc
te prius sit an invoca
nesciens te? Aliud er
nesciens. An potius
Quomodo autem inv
crediderunt? Aut qu
praedicante? Et laud
requirunt eum. Qua
et invenientes laudab
domine, invocans te,
praedicatus enim es

Figure 3: EVT 2.0 - Critical edition. Collation of witnesses

The poster that we would like to present at the DH 2016 will show how far EVT has progressed and matured in terms of features, code robustness and UI innovation especially with regard to critical edition support and handling of manuscript variants.

During the poster session you will be able to try a working demo of the software.

Bibliography

Angular.js. https://angularjs.org/ (accessed 28 February 2016)

Codice Pelavicino digital edition.http://labcd.humnet.unipi.it/ (accessed 28 February 2016)

CVCE (Centre Virtuel de la Connaissance sur l'Europe).http://www.cvce.eu/ (accessed 28 February 2016)

Digital Vercelli Book. *Beta edition*. http://vbd.humnet.unipi.it/beta/ (accessed 28 February 2016)

EVT. *Home page*, https://sourceforge.net/projects/evt-project/ (accessed 28 February 2016)

Hugh Cayless. *Representing Textual Variation (2nd Draft)*. https://docs.google.com/document/d/10R5FfpvCh9v2c2zeG1hgYMcyuT80-PfHiaWYVrLf56k (accessed 28 February 2016)

TEI Guidelines.*Chapter 12 Critical Apparatus.*http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html (accessed 28 February 2016)

TEI Wiki.*Critical Apparatus Workgroup*. http://wiki.tei-c.org/index.php/Critical_Apparatus_Workgroup (accessed 28 February 2016)

TEI Wiki.*TEI Special Interest Group on Manuscripts* (TEI MS SIG). http://wiki.tei-c.org/index.php/SIG:MSS (accessed 28 February 2016)

Notes

[1] Edition Visualization Technology: https://sourceforge.net/projects/evt-project/

[2] Digital Vercelli Book beta edition: http://vbd.humnet.unipi.it/beta/

[3] Codice Pelavicino digital edition: http://labcd.humnet.unipi.it/

[4] http://www.cvce.eu/

[5] Especially if compared to the first version presented at DH 2014.

[6] AngularJS: https://angularjs.org/

[7] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html

[8] For more information see: *TEI Wiki. Critical Apparatus Workgroup*, *TEI Wiki. TEI Special Interest Group on Manuscripts* and *Hugh Cayless. Representing Textual Variation*

Testing Delta on Chinese Texts

Keli Du
keli.du@stud-mail.uni-wuerzburg.de
Universität Würzburg, Germany

Delta (Burrows, 2002) is a measure, which has already been proven as a reliable method to resolve authorship attribution problems in different languages such as English and German. However, there has been no report about the accuracy of Delta on Chinese texts so far. As such, I set an experiment to test it. The tests cover both modern and classical Chinese because of the grammatical and lexical differences between them.

First I determined whether Delta works on modern Chinese. After that I did tests on classical Chinese. At last, I tested the *Dream of the Red Chamber* (DRC, 红楼梦)[1]. The number of authors of DRC is a classic question of Chinese literary studies. The tool I used in the experiments is "stylo", an R package introduced in the context of stylometry in 2013 (Eder, 2013). Using "stylo" I have done cluster analysis. All texts of one author should stay in one group. Misplaced texts are considered as mistakes. The more mistakes Delta makes, the less Delta is appropriate for Chinese.

Working on Chinese language processing is different compared to languages like English. The greatest challenge lies that we are unable to recognize the boundary of words because there are no spaces between words. There are two possibilities to address this problem: (i) by using a segmenter to split a text into words and select words as the textual feature, or (ii) by selecting character N-grams as the feature. Both solutions were tested here and the results are presented as a comparison.

For my first experiment I gathered 45 modern Chinese texts from 6 authors. I used the Stanford segmenter to split the texts and select both words and characters as features. The results showed that Delta is reliable (Fig. 1). With the 100 most frequent words bigrams Delta correctly identifies 38 of 45 texts. The best results, 43 of 45 texts, occur with the 200 to 700 most frequent character bigrams or most frequent words unigram.

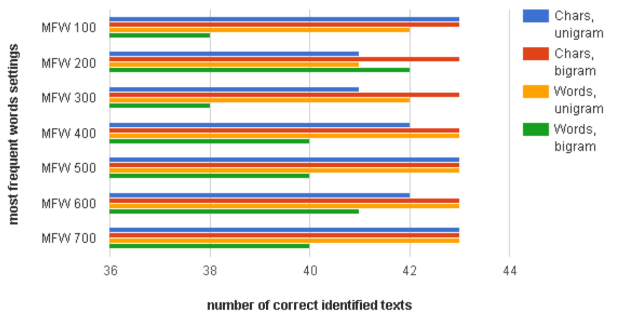


Fig. 1 Delta test results for four sets of features in 45 modern Chinese texts

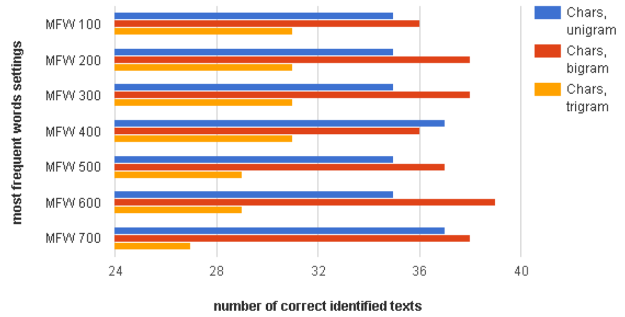


Fig. 2 Delta test results for three sets of features in 40 classical Chinese documents

After the tests on modern Chinese, I proceeded with my second experiment on classical Chinese. I took 4 chapters each randomly from 10 novels from the Ming and Qing Dynasties (16th to 19th century) and built a corpus of 40 documents. One problem was that the Stanford segmenter did not work anymore, because the segmentation standards are not suitable for classical Chinese. Hence the only option was to take characters as feature. The results showed that Delta also works (Fig. 2). While many mistakes occurred with characters trigrams, taking characters bigrams for the tests achieved a high level of accuracy. With 600 most frequent characters 39 of 40 documents were correctly identified.

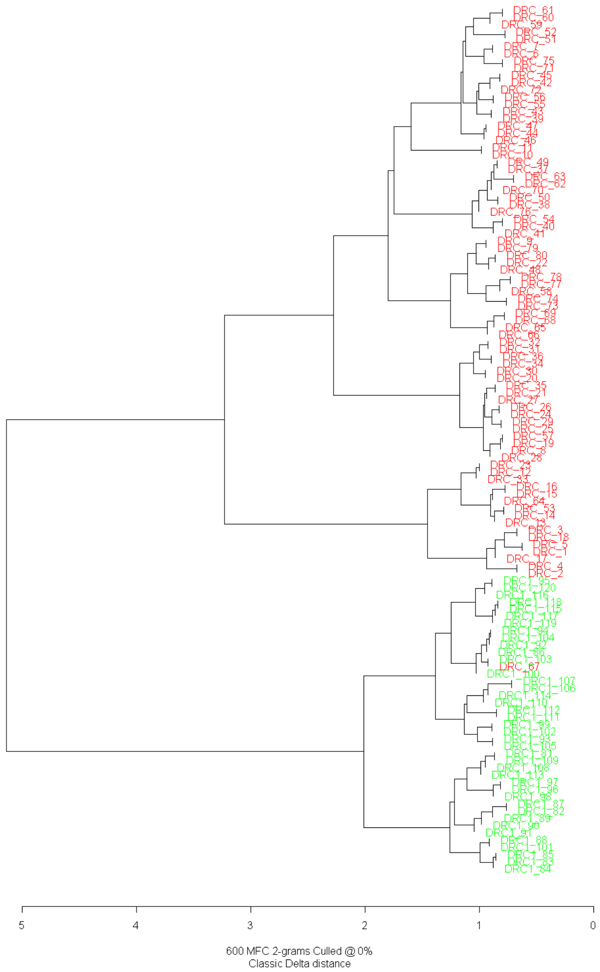


Fig. 3 Delta test results of DRC, (600 MFC, 2-grams)

The first two experiments confirmed Delta as a valid measure for both modern and classical Chinese. In the third experiment Delta was applied to *Dream of Red Chamber* (DRC)[2]. As one of the most famous Chinese classic novels, DRC was written by Cao Xueqin (曹雪芹) during the 18th century. The first version had 80 chapters. However in 1791 Gao E (高鹗) and Cheng Weiyuan (程伟元) published another edition with 120 chapters. They claimed that theirs was the complete version. Since there, there has been a constant debate about the number of authors of DRC. Some scholars think that Cao penned all 120 chapters. Some beg to differ. According to a study by

Hu Shih (胡适) (Hu, 1998), the first 80 chapters of DRC are the original work of Cao and the last 40 chapters are written by Gao. Hu's research is now widely accepted in China. Some modern research approaches also suggest that the first 80 chapters and the last 40 chapters of DRC are written by two different authors. They also find evidence that Chapters 64 and 67 may not be written by Cao (Hu, 2014; Tu, 2013).

My experiment suggested the same conclusion as the other scholars that DRC is written by two different authors (Fig. 3). The texts were divided into two groups. Red texts represent the first 80 and green texts are the rest 40 chapters. Delta also suggests that Chapter 67 is written by the second author.

Bibliography

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship, *Literary and Linguistic Computing*, **17**(3): 267-87.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools, *Digital Humanities 2013: Conference Abstracts* pp. 487-89.

Hu Shhi (1988). 胡适红楼梦研究论述全编 *[Hu Shihs Analysis of Dream of Red Chamber]*, Shanghai Guji Chubanshe (Shanghai Classics Publishing House).

Hu, X., Wang, Y. and Wu, Q. (2014). Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber, *Advances in Adaptive Data Analysis*, 1450012.

Tu, H. C. and Hsiang, J. (2013). A Text-Mining Approach to the Authorship Attribution Problem of Dream of the Red Chamber, *Digital Humanities 2013: Conference Abstracts*, pp. 441-43.

Notes

[1] I focussed on the Classic Delta in my work. Other variations of Delta like Eder's Delta, Argamon's Linear Delta and so on will not be tested.

[2] According to Tu's paper (2013) the DRC under http://cls.hs.yzu.edu.tw/hlm/read/TEXT/TEXT.ASP is „the closest to the earliest editions", which was taken for my study.

The Categories of Philosophy in the Digital Era

Anthony Durity
anthony@durity.com
University College Cork, Ireland (UCC)

Gabriela Vulcu
gabriela.vulcu@insight-centre.org
Insight Centre for Data Analytics at National University of Ireland, Galway (NUIG)

Georgeta Bordea
georgeta.bordea@insight-centre.org
Insight Centre for Data Analytics at National University of Ireland, Galway (NUIG)

James O'Sullivan
josullivan@psu.edu
Pennsylvania State University, USA (PSU)

Jasenka Eva Jones
92562213@umail.ucc.ie
University College Cork, Ireland (UCC)

Knowledge Representation

The research we present is a digital reworking of the categories. This research forms part of a larger work which explores the impact of computational methods on philosophy. Nearly 40-years-ago Aaron Sloman anticipated this *Computer Revolution in Philosophy* (Sloman, 1978). This revolution is as yet not fully under way, we would argue, at least not to the same extent that it is currently transforming the humanities and has already transformed the sciences. Our research adds momentum.

A distinction must be made between the general notion of category in philosophy and the top-level categories of metaphysics. When we refer to the categories of virtue of vice, for instance, it is to the general notion of category to which we refer. When we refer to the ontologies of Aristotle, Kant, Peirce, Whitehead, or whoever, it is to the top-level categories that we refer. Every object/entity there is, without exception, participates in at least one of these top-level categories because these categories indicate universal properties. Properties such as: quality, substance, modality, form, and so on. C. S. Peirce arguably took these investigations to their logical conclusion with his triad of logical terms: firstness, secondness, thirdness. John F. Sowa in *Knowledge Representation* (Sowa, 2000) presents a historical account of this philosophical activity and its connections with software modelling, we regard our research is a continuation of Sowa's investigations.

Saffron (Bordea, 2013) is a domain modelling research tool that uses novel natural language parsing and taxonomic techniques. Specifically, it performs domain adaptive extraction of topical hierarchies. *Saffron* was conceived and is being developed at the Insight Centre for Data Analytics in the National University of Ireland, Galway.

PACT.x (Durity, 2015) is a purpose-built corpus of philosophical texts, these are stored in plain text with their associated metadata and they reflect 2500 years of mainly Western thought, from Plato's *Symposium* to Wittgenstein's *Tractatus*. *PACT.x* was conceived and is being developed at the Digital Arts and Humanities program in University College Cork, Ireland.

We arrive at a digital reworking of the categories in a number of algorithmic steps.

(1) By performing a semantic-based analysis of the *PACT.x* corpus using *Saffron* we obtain result set α, $R\alpha$, which comprises a sequence of terms, $S\alpha$, and related taxonomy, $T\alpha$, of philosophical concepts (Vulcu, 2015) and their corresponding visualisation - Figure 1. Sequence $S\alpha$ is ordered by a weighted combination of frequency and coherence, meaning frequency of occurrence of a word within the corpus coupled with terms that contain a word from the domain model, and terms that appear in the context of a word from the domain model. The structure and relation of the topics in taxonomy $T\alpha$ are drawn from result set $R\alpha$. The directed edges of the graph are constructed using the broader/narrower than relation from WordNet[‡] and the graph is pruned using the ChuLiu/Edmonds algorithm resulting in a directed acyclic graph.

The first 30 terms in sequence α are:

1 Nature	16 Truth
2 Life	17 Matter
3 God	18 Cause
4 Knowledge	19 Mind
5 Law	20 Terms
6 Words	21 Object
7 Principle	22 Character
8 Sense	23 Soul
9 Power	24 Person
10 World	25 Philosophy
11 Time	26 Moral
12 Reason	27 Pleasure
13 Body	28 Experience
14 Idea	29 Government
15 Form	30 Action

(2) We Iterate through each of the terms in result set α in turn and query the open knowledge graph Wikidata[1] to perform further analysis. By tracing the compositional relationship properties (instance-of, subclass-of, part-of) of each term we infer the compositional structure of these concepts and obtain ultimately another sequence, $S\beta$, and related taxonomy, $T\beta$. The list of topics that is sequence $S\beta$ is more or less identical to sequence $S\alpha$ save for disambiguation and completeness. Note that taxonomy $T\beta$ differs from taxonomy $T\alpha$ in that it reflects the mereological structure of Wikidata rather than WordNet.

(3) Taxonomy $T\beta$ is then divided into two parts using a heuristic method: the base taxonomy, $T\gamma$, which comprises the most abstract part, and superstructure which comprises the rest.

Taxonomy $T\gamma$ is the digital reworking of the categories.

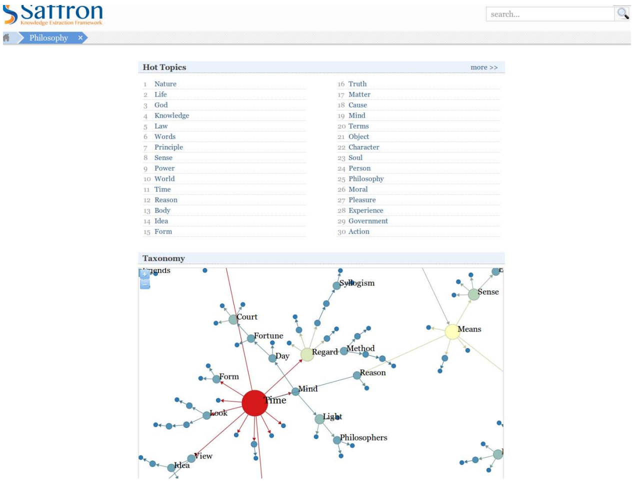
Critical Assessment in Context

We compare taxonomy γ with the traditional categories of philosophy to see how the singular vision of individuals (Aristotle, 1938; Kant, 1787, for instance) differs from what may be viewed as a complex adaptive system (WordNet and Wikidata). To clarify, the idea here is to contrast the semi-isolated contemplation of individuals from different eras with the distillation of the organic collective behaviour of billions of wiki edits. In what way does a digitally generated ontology differ from the standard hand-built ones? For a contrasting approach see the recent theoretical work *Foundations of an Ontology of Philosophy* (Grenon and Smith, 2011) with attendant Web Ontology Language applied work *Philosophy Ontology* (Grenon, 2007)

We validate resultset α against authoritative philosophical reference works such as *The Oxford Dictionary of Philosophy* (Blackburn, 2008), *Routledge Encyclopedia of Philosophy* (Craig, 1996), and so on. In this way we are able to understand how a distant reading (to use a term that has almost become a genericized coinage) of topics by machine algorithms compares to professional reference works. Is it possible that topics that philosophers have touched on down through the ages, as highlighted by the machine, haven't gotten the attention they deserve? If so, why? Is it possible that there are topics which are considered important enough to gain a reference entry that have been overlooked by the machine? Again, what could be the reason? Because the topics in resultset α are ordered by frequency we now have a rough quantitative measure of the relative usage of different philosophical concepts which perhaps can be used as a proxy for their relative human significance; reference works obviously arrange entries in alphabetical order having no other metric.

Our research indicates that novel computational methods may be brought to bear on the task of philosophy. It suggests a radically different route than that taken by others, such as that of modelling the formal systems discussed within philosophical works, as exemplified by *The Philosophical Computer* (Grim, Mar, and St. Denis, 1998). It suggests also that we should take a sustained look beyond the more, if we may say, pedestrian use of the computer as an educational and communication tool as discussed, for example, in *Cyberphilosophy* (Moor and Bynum, 2002).

Though some way off from becoming part of the standard philosophical curriculum the investigations we have undertaken indicate further avenues of fruitful research and hint at future novel results once the methods have been refined through third-party experimentation, replication, and feedback.



Hot Topics more >>

1 Nature
2 Life
3 God
4 Knowledge
5 Law
6 Words
7 Principle
8 Sense
9 Power
10 World
11 Time
12 Reason
13 Body
14 Idea
15 Form
16 Truth
17 Matter
18 Cause
19 Mind
20 Terms
21 Object
22 Character
23 Soul
24 Person
25 Philosophy
26 Moral
27 Pleasure
28 Experience
29 Government
30 Action

Taxonomy

Figure 1: Visualisation of philosophical topic list and taxonomy

Bibliography

Aristotle. (1938). *The Organon*. In H. P. Cooke, H. Tredennick, and E. S. Forster, (Eds.). Cambridge, Mass.; London: Harvard University press; W. Heinemann Ltd.

Blackburn, S. (Ed.). (2008). *The Oxford Dictionary of Philosophy* (2nd ed.). Oxford University Press.

Bordea, G. (2013). *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. PhD dissertation. Retrieved from http://aran.library.nuigalway.ie/xmlui/handle/10379/4484

Craig, E. (Ed.). (1996). *Routledge Encyclopedia of Philosophy*. Routledge. Retrieved 5 March 2016, from https://www.rep.routledge.com/

Durity, A. (2015). Philosophy Archive of Clear Text, *PACT.x*.
Corpus. http://dh.ucc.ie/corpus-builder/texts.xml **Grenon, P.** (2007). Philosophy Ontology. *Web Ontology Language File*. (5 March 2016)http://ontology.buffalo.edu/philosophome/pdcphilontology-v1.owl

Grenon, P., and Smith, B. (2011). Foundations of an Ontology of Philosophy. *Synthese*, **182**(2): 185–204.

Grim, P., Mar, G., and St. Denis, P. (1998). *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*. Cambridge, Mass.: MIT Press.

Kant, I. (1787). *The Critique of Pure Reason*. In M. D. Meiklejohn, Trans, Raleigh, N.C.: Alex Catalogue.

Moor, J. H., and Bynum, T. W. (2002). *Cyberphilosophy: The Intersection of Philosophy and Computing*. Malden, MA: John Wiley and Sons.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, Science, and Models of Mind*. Atlantic Highlands, N.J.: Humanities Press.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove: Brooks/Cole.

Vulcu, G. (2015). Philosophy Texts - One Word. (31 October 2015) http://140.203.155.226:8000/philosophy_one_word/

Notes

[1] https://wordnet.princeton.edu/ "WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations."

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page "Wikidata is a free linked database that can be read and edited by both humans and machines." "Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects[…]" Wikidata is a triplestore. A triplestore is a database of semantic web triples, each with the form: subject, predicate, object. It is no coincidence, we suggest, that the ultimate store of structured data that has evolved on the web shares the same form as Peirce's triad of logical terms.

"Remapping Leigh Hunt's Circles": Voyant Tools and Hunt's Dramatic Criticism

Michael Eberle-Sinatra
michael.eberle.sinatra@umontreal.ca
Université de Montréal, Canada

Stéfan Sinclair
stefan.sinclair@mcgill.ca
McGill University, Canada

Emmanuel Château
emmanuel.chateau.dutier@umontreal.ca
Université de Montréal, Canada

"Remapping Leigh Hunt's Circles" is an ambitious project that explores Leigh Hunt's central position in the London literary and critical scene of the first half of the nineteenth century, through the lens of digital humanities tools. Hunt is today considered one of the key figures of the Romantic period in England, known for his work as editor, journalist, poet, and facilitator. Numerous articles, essay collections, biographies, and monographs published in the last fifteen years have made this clear. Hunt's contribution to Romantic and Victorian literature was as extensive as it has proven durable, in matters as various as prosodic experimentation and the modernization of the magazine essay. Yet little work (beyond some biographical notes) has been done on the second half of his life, a period that was as productive as the first, and during which Hunt was intimate with many of the finest writers of the time, and continued to contribute to London's literary circles through the ongoing publication of critical essays in periodicals and anthologies. This project aims to redress this imbalance/oversight and reassert Hunt's place in the Romantic and Victorian eras, as well as his continuing significance for understanding the London literary scene between 1805 (publication of his first critical essay) and 1859 (date of his death, with his last article published only a few weeks before).

"Remapping Leigh Hunt's Circles" makes a case for Hunt's position as a key critical voice in London beyond his already established prominence during the *Examiner* years. It does so through a careful analysis of his critical reviews and essays (with a specific focus on his drama criticism to underscore Hunt's ongoing engagement with the public sphere) published during his entire career, which spanned the first half of the nineteenth century. Data mining and textual analysis offer exciting opportunities to bring together different sets of data which, when prepared to the highest standard of text encoding, can yield new and innovative results that encourage reconsideration of preconceived notions regarding the transfer of ideas from one author to another, or one literary genre to another. The results of the research undertaken in "Remapping Leigh Hunt's Circles" will be presented in a collaborative, visual context that reimagines the digital scholarly edition as a transparent workspace in which established primary objects from existing databases can be gathered, organized, correlated, annotated, and augmented by multiple users in a dynamic environment.

All the texts prepared for inclusion in our project are encoded to the Text-Encoding-Initiative (TEI) standards. The mark-up language and quality controls for improving metadata in all the resources provide more accurate search and discovery, allow for the presentation of well-supported content on multiple devices and develop tools for assembling, archiving and indexing research objects and artifacts. Ongoing work on this platform will enable researchers to undertake world-class research by providing the means to link data-sets to published content, encouraging data reanalysis, replication studies, and data re-purposing, all of which improve research quality and efficiency.

Our poster will report on the first year of this project, and the implementation of the latest version of the *Voyant Tools* to examine the dramatic essays written by Hunt between 1805 and 1813 (when he was sentenced to two years in prison for libel against the Prince Regent). We will showcase in particular two aspects of the integration between the Hunt archives and Voyant Tools. First, the ability to identify and visualize named entity connections and their networks across multiple documents (this a refinement of the previous RezoViz tool in Voyant). The Hunt collection presents an ideal corpus for network exploration given the interconnectedness of the people, locations and events that animate the documents. Second, Voyant provides a generic and customizable way of presenting a web-based corpus catalogue with the same kinds of faceted browsing and advanced querying capabilities we have come to expect from library databases and online stores. A further benefit of this functionality is the ability to create dynamic subsets of a corpus to examine more closely (in other words, using a catalogue skin in Voyant to create worksets destined for Voyant's more conventional analytic skin).

The "Remapping Leigh Hunt's Circles" is essentially a project of digital text editing and literary criticism whereas Voyant Tools is essentially a software platform for reading, analyzing and visualizing digital texts. These are separate traditions and separate concerns, but this poster will demonstrate the value of symbiotic development: both projects benefit from the collaboration.

Bibliography

McGann, J. (2014). *A New Republic of Letters*. Cambridge, MA: Harvard UP.

Sinatra, M., and Sinclair, S. (2015). Special issue "Repenser le numérique au 21ème siècle". *Sens public* (hiver 2015).

Sinatra, M. (2015). "Representing Leigh Hunt's Autobiography". *Virtual Victorians: Networks, Connections, Technologies*. Eds. Stauffer, A. and Alfano, V. R., Palgrave.

Sinclair, S., Rucker, S. and Radzikowska, M. (2011). *Visual Interface Design for Digital Cultural Heritage*. Ashgate.

The Digital Émigré: Russian Periodical Studies and DH in the Slavic Fields

Natalia Ermolaev
nataliae@princeton.edu
Princeton University, USA

Philip Gleissner
pg4@princeton.edu
Princeton University, USA

Digital Humanities has seen slow adoption in the Slavic language and literature fields in North American academia. This issue frames our project, the Digital Émigré, a digital resource for exploring Russian émigré periodical literature. Our project has a threefold aim. As periodical studies scholars, we want to enable access to Russian émigré journals for new audiences. As digital humanists, we believe that DH tools and methodologies can facilitate new forms of knowledge about twentieth-century Russian, and more broadly diaspora, literary and cultural history.

Finally, as Slavists, we hope our project will be a hub for discussion about the applicability of DH theory and practice for scholars working with Russian-language material.

At this pilot stage, Digital Émigré is a web-based searchable database of article-level metadata of Russian-language journals published outside of Russia in the twentieth century. Our pilot contains four titles (approximately 100 issues and 1,500 articles): *Novoselye* and *Novyi zhurnal* were published in the 1940s in New York, and *Sintaksis* and *Kontinent*, in the late 1970s and 1980s in Paris. Our pilot site provides insight into literary culture at both the beginning and end of the Cold War, bookending the twentieth-century Russian diaspora experience. Digital Émigré is intended to scale, and will eventually contain additional titles and new functionality.

We will highlight the main scholarly avenues that DH methods allow us investigate, such as mapping networks of co-publication, tracking evolving political, social and cultural concerns of émigrés over the course of the Cold War, demonstrating the increased opportunities for émigré women as editors and contributors, and highlighting the proportion of original vs. re-printed work in émigré publications. This way, our project encourages experimentation that will enrich the study of Slavic periodical culture: accessing journals through their data can challenge narratives that are often framed by retroactive canonization, close reading and focus on individual authors. Digital Émigré thereby bridges philological approaches and sociological questions about intellectual networks and communities of artistic production.

The poster address the project's core technical design: our strategy for data modeling and management and database design. We will also present our plans for next steps, which is to provide full-text access and to federate our titles with other digital periodical collections. For this, we are designing a TEI schema modeled on major periodical studies digital collections - specifically the Blue Mountain Project at Princeton University (http://bluemountain.princeton.edu and the Yellow 90's Online at Ryerson University (http://www.1890s.ca).

We will also discuss the specific challenges of working with Russian language material and Cyrillic script, such as character encoding, transliteration, translation, and tokenizing and stemming. These issues can be barriers to success when working with popular DH tools that are developed primarily for Western scripts and languages, and we will show our solutions for using some well-known tools for: data normalization (OpenRefine), text analysis (Voyant), network analysis (Gephi), visualization (Raw, Palladio), and topic modeling (MALLET).

Digital Émigré is committed not only to the exploration of the intellectual experience of diaspora cultural life. As a digital humanities project, it is itself invested in building intellectual communities around the engagement with this material and its afterlife. It aims to foster contact between scholars working with Russian and other Slavic languages internationally, especially through the discussion of issues of interoperability and creating multilingual digital research environments.

Faraway, So Close!: Reading Adeline Mowbray Closely Using Topic Modelling

Michael Gregory Falk
michaelgfalk@gmail.com
University of Kent, UK, United Kingdom

1. Introduction: The "reading" debate

Digital humanists disagree fervently about the nature of reading, and how computers can change the way we do it. Some advocate "distant reading" as a radically new form of inquiry (Burdick et al., 2012, Jockers, 2013, Moretti, 2013). Others argue that computers can improve and enrich traditional modes of reading (Burrows, 1987, McGann, 2001, Ramsay, 2005, Ramsay, 2011). McCarty subsumes this debate in his third way of "interactive modelling" (2005). Pasanek, meanwhile, offers a whimsical alternative with his "desultory reading" (2015).

The hot polemic of this debate obscures a fundamental point: all these kinds of reading are deeply intertwined, and none should be treated as an exclusive option.

To demonstrate this, I apply a classic *distant* reading technique, topic modelling, to a corpus of texts, in order to assist in my *close* reading of a single novel, Amelia Opie's *Adeline Mowbray* (1804). This is a good novel to test new methods on, because it poses stark interpretive problems. Its heroine is a radical who dies repenting her earlier beliefs: are we supposed to admire or condemn her? For two centuries, readers have disagreed, some finding the novel conservative (Tomalin, 1974; Butler, 1987), others radical (Kelly, 1981; King, 2009)—its original readers were mostly just confused (Cooper, 2001). If digital reading can help us answer this question, then we will learn more about the nature and use of digital reading itself.

2. Methods

2.1. Software. I used the popular MALLET package to perform Latent Dirichlet Allocation on *Adeline Mowbray* and 55 other realist novels from the period 1776-1822 (Mccallum, 2002). Blei and Jockers describe the method comprehensively (Blei, 2012, Jockers, 2014).

2.2. Corpus construction. A comprehensive selection of similar contemporary novels was taken from high-quality online archives.

2.3. Data preparation. Scholars disagree about how texts should be prepared for modelling: should they be chunked by paragraph (Algee-Hewitt et al., 2015), or by *x*-length chunk (Jockers, 2013, Jockers and Mimno, 2013, Jockers, 2014)? *X*-length chunks make it easier to include dialogue in the analysis, so these were preferred. The length was set at 125 words, any longer made close reading of the chunks harder; any shorter made the "topics" incoherent.

2.4. Parameters. I used hyperparameter optimization and Jockers' stopword list; excluded characters' names from the analysis; and set the number of topics at 150, avoiding the problems of incoherent and "chimera" topics identified by Schmidt (2012).

3. Discussion and Results

Applying this technique to the novels in my corpus produced two main kinds of results. First, the model identifies key "topics" in the corpus:

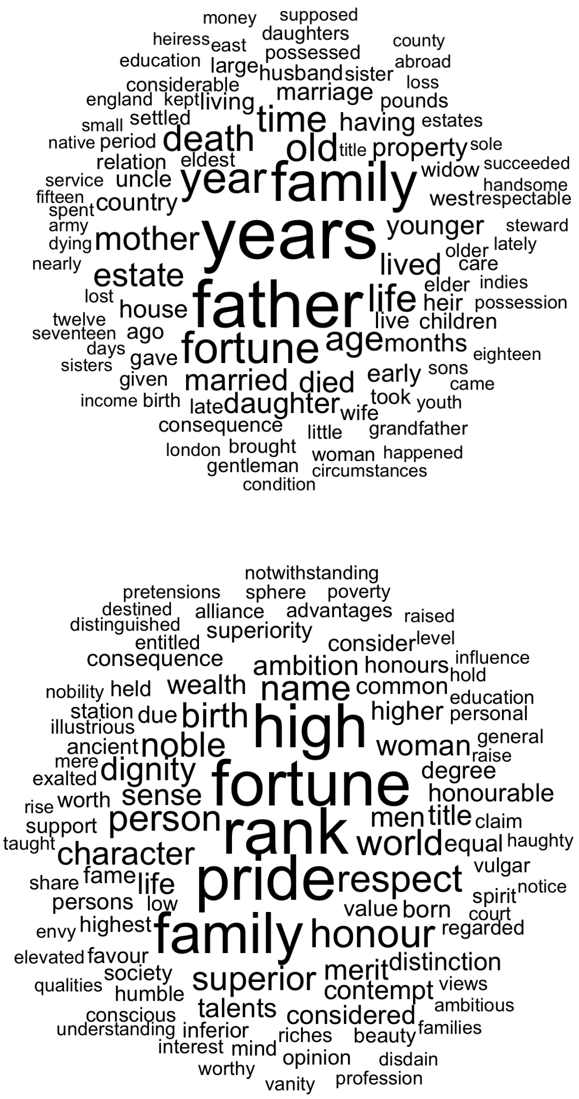
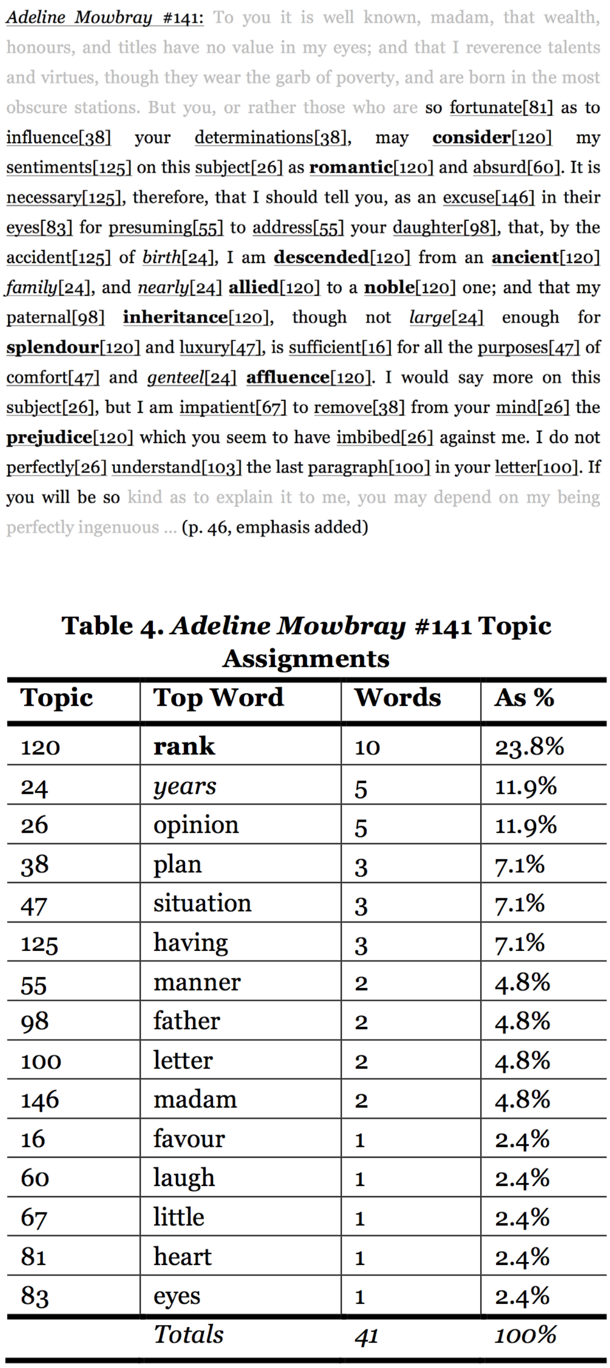


Figure 1. Topics 120 and 24

Such topics uncover hidden patterns between words in the corpus: for instance, the prominence of the words "year" and "years" in the discussion of a family's marital history in novels of this period.

Secondly, as Rhody (2012) has shown, topic modelling enables us to visualise the linguistic composition of a passage:



Adeline Mowbray #141: To you it is well known, madam, that wealth, honours, and titles have no value in my eyes; and that I reverence talents and virtues, though they wear the garb of poverty, and are born in the most obscure stations. But you, or rather those who are so fortunate[81] as to influence[38] your determinations[38], may **consider**[120] my sentiments[125] on this subject[26] as **romantic**[120] and absurd[60]. It is necessary[125], therefore, that I should tell you, as an excuse[146] in their eyes[83] for presuming[55] to address[55] your daughter[98], that, by the accident[125] of *birth*[24], I am **descended**[120] from an **ancient**[120] *family*[24], and *nearly*[24] **allied**[120] to a **noble**[120] one; and that my paternal[98] **inheritance**[120], though not *large*[24] enough for **splendour**[120] and luxury[47], is sufficient[16] for all the purposes[47] of comfort[47] and *genteel*[24] **affluence**[120]. I would say more on this subject[26], but I am impatient[67] to remove[38] from your mind[26] the **prejudice**[120] which you seem to have imbibed[26] against me. I do not perfectly[26] understand[103] the last paragraph[100] in your letter[100]. If you will be so kind as to explain it to me, you may depend on my being perfectly ingenuous ... (p. 46, emphasis added)

Table 4. *Adeline Mowbray* #141 Topic Assignments

Topic	Top Word	Words	As %
120	**rank**	10	23.8%
24	*years*	5	11.9%
26	opinion	5	11.9%
38	plan	3	7.1%
47	situation	3	7.1%
125	having	3	7.1%
55	manner	2	4.8%
98	father	2	4.8%
100	letter	2	4.8%
146	madam	2	4.8%
16	favour	1	2.4%
60	laugh	1	2.4%
67	little	1	2.4%
81	heart	1	2.4%
83	eyes	1	2.4%
	Totals	*41*	*100%*

Figure 2. The topic composition of a single fragment

Such visualisation is an invaluable tool of discovery, directing the scholar's eye to significant patterns in a text's language. In this case, the prominence of topics 120 and 24 in the passage turned out to be crucial.

In this passage, Glenmurray uses these topics to persuade Adeline's mother that he is a good suitor. He is a radical who has published philosophical tracts against marriage. If the novel presents him as the ideal suitor for Adeline's hand, we may assume that the novel itself has radical sympathies. As close reading reveals, topics 120 and 24 are often used to describe past or potential suitors, in both this and other novels from the corpus. In *Adeline Mowbray*, the language of these topics is used to describe many characters negatively, and only one positively—Glenmurray. This is compelling new evidence that he is the ideal suitor, and that the novel therefore portrays him and his radicalism sympathetically.

This reading was both "close" and "distant." The reading was "distant" because it involved statistical analysis of 65,000 125-word chunks. But these patterns observed from a distance were used to unravel the complexities of a closely-read text. Close reading reveals strengths and flaws of the distant reading tool: it turns out that LDA assigns words to topics differently each time it is run, and is scuppered by novelists who write idiomatically. Close and distant reading, it seems, are so closely bound together that they are part of a single process. Their messy, unpredictable connections suggest a new kind of readerly ideal for the Digital Humanities: anarchic reading.

Bibliography

Algee-Hewitt, M., Heuser, R. and Moretti, F. (2015). Paragraphs: The Forgotten Middle. *Micromégas: The Very Small, the Very Large, and The Object of Digital Humanities*. Stanford, CA.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, **55**: 77-84.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. S. and Schnapp, J. (2012). *Digital_Humanities*. Cambridge, MA and London, MIT Press.

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Criticism*. Oxford, Clarendon Press.

Butler, M. (1987). *Jane Austen and the War of Ideas*. Oxford, Clarendon Press.

Cooper, C. M. (2001). Reading otherwise: The abortive politics of Adeline Mowbray, or the mother and daughter. *European Romantic Review*, **12**: 1-42.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago and Springfield, University of Illinois Press.

Jockers, M. (2014). *Text Analysis with R for Students of Literature*. Cham, Springer.

Jockers, M. L. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, **41**: 750-769.

Kelly, G. (1981). Amelia Opie, Lady Caroline Lamb, and Maria Edgeworth: Official and Unofficial Ideology. *Ariel. A Review of International English Literature Calgary*, **12**: 3-24.

King, S. (2009). The 'Double Sense' of Honor: Revising Gendered Social Codes in Amelia Opie's Adeline Mowbray. In: Wallace, M. L. (ed.), *Enlightening Romanticism, Romancing the Enlightenment: British Novels from 1750 to 1832*. Surrey, England: Ashgate.

McCallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit*.

McCarty, W. (2005). *Humanities Computing*. Houndmills, Palgrave.

McGann, J. J. (2001). *Radiant Textuality: Literature After the World Wide Web*. Houndmills, Palgrave.

Moretti, F. (2013). *Distant Reading*. London and New York, Verso.

Pasanek, B. (2015). *Metaphors of Mind: An Eighteenth-Century Dictionary*. Baltimore, Johns Hopkins University Press.

Ramsay, S. (2005). In Praise of Pattern. *TEXT Technology*, **2**: 177-90.

Ramsay, S. (2011). *Reading Machines: Towards and Algorithmic Criticism*. Urbana, Chicago and Springfield, University of Illinois Press.

Rhody, L. M. (2012). Topic Modelling and Figurative Language. *Journal of Digital Humanities* [Online], 2. Available: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities. Journal of Digital Humanities [Online], 2. Available: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

Tomalin, C. (1974). *The Life and Death of Mary Wollstonecraft*. London, Weidenfeld and Nicholson.

The Corpus of Revenge Tragedy (CoRT): Toward Interdisciplining Early Modern Digital Humanities and Genre Analysis

Danielle Marie Farrar
dmfarrar@mail.usf.edu
University of South Florida, United States of America

Discoveries in anatomy and dissection in early modern England and the lively publishing economy of medical texts during this period undoubtedly played a critical role in the popularization of medical and anatomical language in early modern English drama. England's mid-sixteenth century saw the beginnings of the culture of dissection (Sawday 1995), and it was not long until language once exclusive to physicians and anatomists found its way onto the early modern stage and invested the tumultuous worlds of revenge tragedy. Just as the human envelope was peeled away on anatomists' tables, in artists' engravings, and by writers' quills in order to permit autoptic access to what was previously veiled, so can Digital Humanities (DH) techniques and processes of inquiry offer new modes of dissection on dramatic texts. The value of DH to and in early modern dramatic studies rests not only in its capacity to complement existing modes of literary analysis but also in its disruption of reading, interpretive, and knowledge construction practices of early English texts.

Additionally, within the work of genre analysis, DH approaches perform a valuable role in the furthered exploration and understanding of a genre's discursivity and position as a cultural phenomenon rather than just a material object (Steggle 2015). In the continued study of early modern revenge tragedy--a genre that has received relatively little attention in comparison to others--DH can prove invaluable by demonstrating how this genre increasingly participated in discourses of medicine and anatomy across the Elizabethan (1558-1603), Jacobean (1603-1625), and Caroline (1625-1642) periods via mapping the movement of anatomical language from the medical register to a literary and dramatic one. As has been rightly suggested, DH significantly expands the purview of what can be questioned by literary analysis to include not just a consideration of literary texts but also larger modes of cultural production (Wilkens 2015). An examination of how revenge tragedies adopted and exploited the medical register is significant to understanding not only the cultural significance of the revenge tragedy genre but also discovering the early modern conceptions of embodiment and the pervasiveness of the contemporary medical arena in popular culture.

This project thus reports on a hybrid corpus linguistics and geohumanities approach to a genre analysis of early modern revenge tragedy, which serves to widen the scope of inquiry and interpretation in respect to the role of the body and anatomical language in revenge tragedy. As a means to map the increasing frequency of anatomical language and medical vernacular in revenge tragedy, a corpus analysis was conducted using AntConc, a digital tool that enables textual analysis and corpora comparison. I created the Corpus of Revenge Tragedy (CoRT) as an experimental corpus of approximately 40 revenge tragedies to compare against other corpora, including Shakespeare's corpus and other larger corpora, such as the Early Modern English Medical Texts (EMEMT) Corpus, the Corpus of English Dialogues (CED), and the Early English Books Online-Text Creation Partnership (EEBO-TCP) Phase I database of more than 25,000 texts. As a means to focus interpretation, I developed a variety of word lists by collating—through traditional concordance-making methods—an Anatomical Lexicon (AL) of 209 representative material and metaphorical words from revenge tragedies. The AL provides first recorded usages of each word, root language(s), as well as select definitions from the *Oxford English Dictionary* and definitions (when available) from Henry Cockeram's *The English Dictionarie* (1623). In a subset of revenge tragedies, I manually flagged all anatomically inflected words. I then broke the list up into 2 primary categories and 13 secondary categories. Anatomically direct words are words that deal *directly* with anatomy (e.g., *eye*, *brain*, *arm*, etc.). Anatomically indirect words are words that deal *indirectly* with anatomy (e.g., *sconce*, *aspect*, *soul*, etc.). The categories are compositely based on the early modern medical texts of Thomas Vicary, John Banister, Andreas Vesalius via Thomas Geminus, Helkiah Crooke, and Ambroise Paré via Thomas Johnson. As such, data mining is an essential method to this project since this is a method that necessitates and encourages interdisciplinary work (Hagood 2012). Data mining has also helped shaped this project both instructively and interpretively since--as DH typically does--it allows not only new modes of investigation but also produces new systems of knowledge making.

As this diachronic project's interest is grounded in a discussion of how early modernity conceptualized not only the body itself but also understood embodiment (and how those two things manifested in revenge tragedy), this project also looks to geohumanities--a branch of DH--as a means to consider the geography of anatomical language and printing trends in early modern London. Through a lexical mapping of anatomical words in conjunction with a mapping of publication locations that printed revenge tragedy texts, we can gain an understanding of the cultural geography of London in respect to the pervasiveness of the medical register. Through various DH methods, the CoRT and AL complement and enrich existing bodies of knowledge about histories of anatomy and medicine, early modern drama, and early modern theories of embodiment.

Bibliography

Hagood, J. (April/May 2012). A Brief Introduction to Data Mining Projects in the Humanities. *Bulletin: American Society for Information Science and Technology*. *https://www.asis.org/Bulletin/Apr-12/AprMay12_Hagood.html*

Sawday, J. (1995). *The Body Emblazoned*. London: Routledge.

Steggle, M. (2008). "Knowledge will be multiplied": Digital Literary Studies and Early Modern Literature. In Schreibman, S. and Siemens, R. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell. *http://www.digitalhumantites.org/companionDLS/*

Wilkens, M. (2015). Digital Humanities and Its Application in the Study of Literature and Culture. *Comparative Literature*. Special section, "Empirical and Systemic Approaches to the Study of Literature and Culture" (accessed 13 March 2016).

3D Scanning for Preservation: Difficulties and Dissemination

Graham Fereday
G.N.Fereday@exeter.ac.uk
University of Exeter, UK, United Kingdom

Michael Mullins
M.K.Mullins@exeter.ac.uk
University of Exeter, UK, United Kingdom

Richard Webb
R.Webb2@exeter.ac.uk
University of Exeter, UK, United Kingdom

Gary Stringer
G.B.Stringer@exeter.ac.uk
University of Exeter, UK, United Kingdom

The use of 3D scanning & visualisation technologies for creation of digital surrogates has begun to gain wider acceptance in the museums and archives sector, largely due to significant advances in technology and price over the last few years. However, the technology is still evolving, and in some situations can struggle to yield accurate results. This poster will look at the pragmatic aspects of scanning for preservation, and will, it is hoped, generate public discussion and sharing of best practice.

We present a number of case studies. Firstly, looking at reflectivity, we show techniques in scanning metallic objects, and non-intrusive ways of reducing and eliminating surface reflection which can introduce unwanted noise and confuse tracking calculations in hand-held scanners.

Secondly, we look at results from a number of objects with significant amounts of very fine surface detail or flexible materials. Various materials such as modern fabrics, leather, thread and wool present particular challenges, and the results of working with these objects will provide the basis of guidance in preparation for undergraduate and postgraduate users of equipment at Exeter, which will be shared publically.

Thirdly, we examine experiments in the use of 3D scanning to evaluate and preserve archaeological finds in situ during excavation, and in particular, we evaluate the accuracy of measurement using hand-held scanners. This has clear advantages if sufficiently reliable, including the taking of otherwise impossible measurement through objects and evaluation through reconstruction of artefact fragments.

Finally, we look at means of disseminating the results of the scanning process, looking at the problems of reducing polygon counts, and the importance of surface texture and overlays in recreating the 'look and feel' of objects. We briefly look at the 'impact' of virtual 3D surrogates against 2D images, in particular referencing a student-engagement project involving the evoking of memory in dementia sufferers, using archival material and 3D visualisations of football memorabilia.

We examine the cross-generational possibilities opened up by engaging groups of students and seniors in the capture process, and suggest further possibilities for triggering memory or emotional responses through the use of detailed and immersive surrogates.

It is highly likely that further case studies will be examined; by June 2016, projects looking at difficult materials such as beeswax and ancient fabric remnants are likely to be under way, and preliminary findings will be included in the poster.

It is hoped that this poster will allow the sharing of ideas and experience both directly at the interactive poster session, and through the linked web objects, where comments can be left and dialogue can continue.

Bibliography

Ioannides, M., Quak, E. (2014). *3D Research Challenges in Cultural Heritage: A Roadmap in Digital Heritage Preservation*. Springer.

Koller, D., Frischer, B., Humphreys, G. (2010). Research Challenges for Digital Archives of 3D Cultural Heritage Models. *J. Comput. Cult. Herit.* 2, 7:1–7:17. doi:10.1145/1658346.1658347

Remondino, F. (2011). Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning. *Remote Sensing* 3, 1104–1138. doi:10.3390/rs3061104

Rodrigues, M.A., Kormann, M., Davison, L., (2011). *A case study of 3D technologies in higher education: Scanning the metalwork collection of museums in Sheffield and its implications to teaching and learning.* Presented at the 2011 International Conference on Information Technology Based Higher Education and Training (ITHET), pp. 1–6. doi:10.1109/ITHET.2011.6018697

Linguistic DNA: Modelling Concepts and Semantic Change in English, 1500-1800

Susan Fitzmaurice
s.fitzmaurice@sheffield.ac.uk
University of Sheffield, United Kingdom

Marc Alexander
marc.alexander@glasgow.ac.uk
University of Glasgow, United Kingdom

Michael Pidd
m.pidd@sheffield.ac.uk
University of Sheffield, United Kingdom

Justyna Robinson
justyna.robinson@sussex.ac.uk
University of Sussex, United Kingdom

Fraser Dallachy
fraser.dallachy@glasgow.ac.uk
University of Glasgow, United Kingdom

Iona Hine
i.hine@sheffield.ac.uk
University of Sheffield, United Kingdom

Seth Mehl
s.mehl@sheffield.ac.uk
University of Sheffield, United Kingdom

Brian Aitken
brian.aitken@glasgow.ac.uk
University of Glasgow, United Kingdom

Matthew Groves
m.groves@sheffield.ac.uk
University of Sheffield, United Kingdom

Katherine Rogers
k.rogers@sheffield.ac.uk
University of Sheffield, United Kingdom

Linguistic DNA is a collaboration by linguists, historians, and digital humanities specialists at the Universities of Sheffield, Glasgow and Sussex. Our aim is to explore the emergence and development of semantic concepts as they are realised in historic textual corpora through a combination of computational processing and data visualisation techniques. This poster outlines the project's goals and methodology, describes the progress made towards those goals, and offers interim results. Prior to the era of big data, semantic research has relied on intuitive selection of concepts worthy for study and has drawn its evidence largely from canonical texts. The advent of large machine-readable textual collections opens the door to new methodologies for research in conceptual history, revolutionising our ability to extract information from such data sets. The Linguistic DNA project explores the use of these techniques for historical semantics, beginning with annotated corpora yet without a predetermined set of concepts to study, the intention being that through text processing and data visualisation, concepts should emerge 'bottom-up' out of collections that extend far beyond the canonical texts.

The main source of data for analysis is the Early English Books Online collection (henceforth EEBO)[1] and Eighteenth Century Collections Online (ECCO).[2] EEBO has been manually transcribed to high accuracy levels by the Text Creation Partnership (TCP) whilst ECCO is partly manually transcribed and partly OCR'd. All are to be annotated with lemma and part of speech information, although our process takes different inputs, and begins with cleaned text. These collections together consist of English-language material printed between the 15th and 18th centuries. Initial stages of investigation involve the development of a software tool or suite of tools to query the data and provide input for visualisation. The success of the visualisations is then evaluated by the project's team of research associates, investigating patterns which emerge, seeking verification of these patterns through returning to the textual source material, and using the resulting insights as input for iterative improvement of the querying and visualisation processes.

In the first year of the project, development of the query software has begun with assessment of the text for potentially challenging features, such as the difficulty posed by pre-standardisation spelling, inconsistent transcription practices, and atypical syntax. Also essential has been identification of the pre-processing required before the texts are analysed, and investigation of existing software packages that might be adapted and extended to meet our research goals. S-Space and BlackLab are examples of tools which might form part of a new pipeline, the components and algorithms of which will be developed by iterative experimentation. The processor will take account of different statistical measures starting with Pointwise Mutual Information, collecting data for a range of proximity windows to assess semantic relevance through distributional semantics techniques. Groups of words with strong patterns of association are output, which are then investigated as candidate concepts. In later stages the project will also use versions of the textual data annotated with sense codes based on the *Historical Thesaurus of English.*[3] This facilitates disambiguation of the senses of homographs, as well as offering another means of assessing relationships between words. To maximise the 'bottom-up' approach to data analysis, the LDNA processor initially indexes and runs queries on every word in the corpus. This avoids presupposing concepts or key terms *a priori*.

Further evaluation will be conducted through the lens of three 'research themes'. Research Theme 1, led by Professor Susan Fitzmaurice at the University of Sheffield, will contextualise the emergence and development of concepts within the historical situations which have instigated and shaped them. Research Theme 2, led by Dr Justyna Robinson at the University of Sussex, will investigate where the boundaries of concepts lie and the families of words which delineate and cross these boundaries. Research theme 3, led by Dr Marc Alexander at the University of Glasgow, will explore moments of rapid change in the lexical items used to instantiate concepts. The research themes begin their work once initial processor development has taken place, and case studies with preliminary findings will be included on the poster. The project runs until 2018.

The Linguistic DNA project is funded by the Arts and Humanities Research Council (project AH/M00614X/1).

Notes

[1] http://www.textcreationpartnership.org/tcp-eebo/

[2] http://gdc.gale.com/products/eighteenth-century-collections-online/

[3] Kay, Christian, Jane Roberts, Michael Samuels, Irené Wotherspoon, and Marc Alexander (eds.). 2015. The Historical Thesaurus of English, version 4.2. Glasgow: University of Glasgow. http://historicalthesaurus.arts.gla.ac.uk/.

The digital breadcrumb trail of Brothers Grimm

Greta Franzini
gfranzini@gcdh.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Emily Franzini
efranzini@gcdh.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Gabriela Rotari
gabriela.rotari@stud.uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Franziska Pannach
franziska.pannach@stud.uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Mahdi Solhdoust
mahdi.solhdoust@stud.uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Marco Büchler
mbuechler@gcdh.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Grimm's *Kinder- und Hausmärchen*: Intratextuality and Intertextuality

Described as 'a great monument to European literature' (David and David, 1964; 180), Jacob and Wilhelm Grimm's masterpiece *Kinder- und Hausmärchen* (hereafter KHM) has captured adult and child imagination for 200 years. International cinema, literature and folklore have borrowed and adapted the Brothers' fairy tales in multifarious ways, inspiring themes and characters in numerous cultures and languages. While commonly and erroneously considered the fathers of the genre, the fairy tales were not original to the Brothers. In fact, Jacob and Wilhelm collected and adapted their stories from earlier works, some of them dating back to as far as the seventh century BC, and made numerous changes to their own collection (David and David, 1964, p. 183), producing seven distinct editions between 1812 and 1857. In these four decades of writing and rewriting, the fairy tales changed in number, style and content in accordance with historical, social and literary influences. And yet, how did forty years of revisions not confuse the tradition and transmission amongst followers and readers? Indeed, some fairy tales were changed almost beyond recognition. What makes them so timeless and memorable? What is it that immortalises these tales? An answer to this question can be found in the *motifs* the Brothers borrowed from earlier traditions and disseminated by way of their famous collection.

Motifs, defined by Prince's *Dictionary of Narratology* (2003) as '[...] minimal thematic unit[s]', pervade the Grimm collection and are stable elements interlacing the seven editions of the KHM. The puss with the boots (*Der gestiefelte Kater*) and the concept of the breadcrumb trail (originating in the *Hänsel und Gretel* tale) are both examples of motifs, and they recur not only throughout the Grimm editions, but also over time and space. The occurrence and repetition of motifs within the Grimm collection is a form of *intratextuality*, a term used to describe the internal relations within a text or an author and, in our case, within the KHM editions. But a motif may also appear in other authors across traditions and languages, thus creating *intertextual* relations, relations that the KHM may have with other texts.

Related work and computational opportunities

The breadcrumb trail generated by these motifs in literary history, and internationally spread through the Brothers Grimm, has been extensively studied by folklorists, historians and literary critics. Akin to memes, motifs are a form of information transfer and reuse, which opens up numerous opportunities for computational research in cultural evolution and transmission. Interestingly, however, the study of motifs has not yet fully explored the affordances of digital methods. Many authoritative volumes and ontologies have been published in print, such as the well-known *Enzyklopädie des Märchens*, or the *Estonian Folktales* and the *Catalogue of Portuguese Folktales*, but only a few digital projects or digital editions of these print sources exist. One such digital initiative is the *Aarne-Thompson's Motif-Index*[1], a crucial contribution to the field, often used as a reference system for the production of folktale catalogues.

The situation, however, is different for fairy tales, inasmuch as digital copies of many folktale collections are freely available from Google Books or the Internet Archive,[2] or from online collections, such as the Nederlandse VolksverhalenBank initiative[3] or the Satorbase project[4], fostering intertextual research never before possible.[5] Indeed, we can now leverage hyperlinks and APIs in order to automatically retrieve specific and previously inaccessible information across the web, and to connect existing resources for comparative studies. Moreover, no effort has yet addressed the cross-cultural relations of fairy tales, giving way to opportunities for interdisciplinary, multilingual and big data research.

Our project

The new project[6] described in this paper is one such opportunity, whereby an international and interdisciplinary team of computer scientists and humanists is semi-automatically crawling digitised texts and the web to produce a multilingual motif index that uses the Grimm collection as its base reference.[7] More specifically, we combine knowledge acquired from existing print and digital resources with the deployment of the Google Search[8] and Google Books APIs[9] in order to automatically retrieve as many motifs across the web in as many languages as possible, and hence to explore the intratextual and intertextual relations that characterise the motifs' hosting texts. The end goal is twofold; on the one hand, we provide a comprehensive reference resource for scholars in the field and interested citizens alike and concurrently revise the Aarne-Thompson Index; on the other, by testing state-of-the-art text reuse and retrieval algorithms on a sample of these diverse and large datasets, we are able to refine our methods in order to accommodate further web-scale queries and thus sharpen our understanding of why and how motifs changed.

Methodology

The case studies we are working with to address our research questions are three Brothers Grimm tales: *Snow White*, *Puss in Boots* and *The Fisherman and his Wife*. These were chosen on the basis of their differing degree of popularity in order to better understand how transmission affects popularisation.

Our research starts with digital and clean copies of the Grimm texts, downloaded and catalogued from TextGrid[10] and Wikisource[11]. Next, our international team of researchers and student assistants collects digitally available translations and/or editions of the three tales in multiple languages[12] and manually enters them into a database, where information about the web source, the tales, the language, the work and the author is stored.[13] Once this manually-compiled dataset is complete, we deploy the TRACER suite of text reuse algorithms (Büchler, 2013) to trace additional motifs in other digital libraries or corpora. At the same time, we use the Google Search and the Google Books APIs to search for motifs at a much larger scale, effectively crawling the web.

Impact

Like the KHM, we believe this project appeals to a wide and diverse audience not only because of its subject matter, but also because of its international and interdisciplinary character. Our international group operates at the intersection of Computer Science and the Humanities in the arena that is Digital Humanities. This project is unique insofar as each and every member of the team can contribute a piece of his or her own culture, adding a personal and familiar touch to this joint endeavour. By exploring these different cultures, we aim to establish fruitful collaborations and, in so doing, broaden the boundaries of the Digital Humanities.

Furthermore, we believe that this project fully engages humanists in the digital process of tracing texts through space and time. Following the motif trail back in time allows humanists to investigate lines of transmission of folktales and to potentially uncover additional trails through which other documents or stories travelled. At the same time, it enables the computer scientists in the team to identify any shortcomings in our algorithms and to better understand what to automatically feature when tracing this type of information in a digital ecosystem.

Bibliography

Büchler, M. (2013). *Informationstechnische Aspekte des Historical Text Re-use*. Ph.D thesis, University of Leipzig.

David, A. and David, M. E. (1964). A Literary Approach to the Brothers Grimm. *Journal of the Folklore Institute*, **1**(3): 180-96. http://www.jstor.org/stable/3813902 (accessed 26 July 2015).

Prince, G. (2003). *Dictionary of Narratology. Revised Edition*. University of Nebraska Press, Lincoln and London.

Notes

[1] *The Aarne-Thompson Motif-Index* can be accessed at: http://www.ruthenia.ru/folklore/thompson/ (accessed 18 October 2015).

[2] For example, the 1550-1553 Venetian collection *Le piacevoli notti* by Giovanni Francesco Straparola, at: https://goo.gl/fAAoJ6 (accessed 21 October 2015).

[3] Available at: http://www.verhalenbank.nl (accessed 1 January 2016).

[4] Available at: http://satorbase.org/ (accessed 1 January 2016).

[5] The increasing availability of digital and digitised assets allows us to access information more easily and to potentially uncover previously unknown or unchartered territory.

[6] Starting in October 2015 and running until December 2018.

[7] The team does not include but consults folklorists. We start with the Grimm collection as we already have clean data to work from.

[8] Available at: https://developers.google.com/custom-search/ (accessed 26 October 2015).

[9] Available at: https://developers.google.com/books/?hl=en (accessed 26 October 2015).

[10] Available at: https://textgridrep.org/browse/-/browse/nxvg_0 (accessed 26 October 2015).

[11] Available at: https://de.wikisource.org/wiki/Kinder-_und_Haus-M%C3%A4rchen_Band_1_%281819%29 (accessed 26 October 2015).

[12] eTRAP is currently a team of twelve people from seven nationalities speaking eleven different languages.

[13] An example may be of use in clarifying the point. Grimm's *Snow White* corresponds to Pushkin's Сказка о Мертвой Царевне и о Семи Богатырях (*The Tale of the Dead Princess and the Seven Knights* in English). The two tales differ in many points, including the title of the tale. In Pushkin the princess is protected by seven *knights* (семь богатырей) whereas in the Grimm tale it is seven *dwarves*. Despite the differences, the motifs of the beautiful princess and of her seven protectors link the two stories. To hyperlink and map these versions and their differences, we use a combination of Thompson identifiers for tales, VIAF identifiers for authors and works, and customised identifiers where existing ones do not apply. This semi-automatic approach allows us to populate our database with both content and metadata, and establish relations between the different versions.

The Southern Netherlands and the Infrastructure of Early Modern Globalisation (1500-1800)

Ulrike Valeria Fuss
ulrike.fuss@arts.kuleuven.be
KU Leuven, Belgium

Christian Pistor
christian.pistor@arts.kuleuven.be
KU Leuven, Belgium

Werner Thomas
werner.thomas@arts.kuleuven.be
KU Leuven, Belgium

Cesar Esponda de la Campa
cesar.esponda@student.Kuleuven.be
KU Leuven, Belgium

Lieve Behiels
lieve.behiels@telenet.be
Lessius Hogeschool, Belgium

Cesar Manrique Figueroa
cesaremanrique@gmail.com
UNAM, Mexico

In the early modern period, books were the most important medium of cultural and scientific exchange. The trade in books and the translation of books were two major aspects in this process. This presentation aims to introduce a research project at the **KU Leuven (Belgium)** that examines the specific contribution of books published in the Southern Netherlands and their role in early modern globalisation. This investigation encompasses four sub-projects, three of which analyse **the importance of books from Southern Netherlands for the cultural life of early modern Spanish America**. These focus on the viceroyalties of New Spain, Peru, New Granada, and Rio de la Plata. The poster introduces the sub-project about the viceroyalty of Peru, thereby outlining the use of digital tools in this research context.

Books from the Southern Netherlands and their role in early modern globalisation (1500-1800)

The fourth sub-project looks at **book translations produced in the Southern Netherlands**. It aims to increase our understanding of the world-system of translations as it existed (and evolved) during the early modern period.

All four sub-projects are unique, in that for the first time, they harness the potential offered by the digitalisation of books and library catalogues allowing us to arrive at more comprehensive conclusions than could be reached in the past.

The two show-cased sub-projects were facilitated by the use of online tools and the effective use of standard software. They are exemplary for new approaches to research questions which could not be posed earlier, because the data was not accessible and/or the amount of resulting detailed information could not be appropriately analysed.

The book trade projects prove the existence of a global network spanning three hundred years, while content analysis based case studies and comparisons made to books by Peruvian authors in the early modern period reveal how publications originating from the Southern Netherlands influenced intellectual elites in the Spanish Colonies. The translation project has documented how translation was used to bolster the image of the Spanish Monarchy and advance the cause of the Counter Reformation.

This presentation gives an **overview of the methodological steps** needed to answer the **research questions** and it presents the **digital tools employed** in these tasks.

The main section provides exemplary representative results based on the evaluation of metadata collected from digital resources such as library catalogues. This data was analysed using standard MS office software.

Initial *results have already been published* in various articles. More comprehensive results are awaiting publication.

Bibliography

Behiels, L.,Thomas, W. and Pistor, C. (2014). Translation as an Instrument of Empire: The Southern Netherlands as a Translation Center of the Spanish Empire, 1500-1700. *Historical Methods*, **47** (3): 113-27.

Pistor, C., Behiels, L. and Thomas, W. (2013). Translation, court networks, and the fashioning of an Imperial image: Charles V and the work of Luis de Ávila y Zúñiga. Bibliothèque d'Humanisme et Renaissance. *Travaux et Documents*, **75** (2): 271-89.

Manrique Figueroa, C. (2013). New Spain's import of Culture from the Southern Netherlands. The case of Books, in: Hyden-Hanscho V. and Pieper R. (Eds.), *Cultural Exchange and Consumption Patterns in the Age of the Enlightenment*, pp. 41-56.

Manrique Figueroa, C. (2013). Studying the book in Hispanic America. The process of consolidation of national identities, *Jaarboek voor Nederlandse Boekgeschiedenis*, pp. 187-200.

Fuss, U. (2012). Books and book trade as trigger of the global modernity, in: *Libro y lectura en la Historia*, EREBEA. Revista de Humanidades y Ciencias Sociales, **2**, (online).

Fuss, U. (2011). From Antwerp to Peru - books from the Southern Netherlands in the 16th century's Viceroyalty, *Jaarboek voor Nederlandse Boekgeschiedenis*, jaargang 18, pp. 115-32.

Manrique F., C. and Thomas, W. (2009). La infraestructura de la globalización: la imprenta flamenca y la construcción del imperio hispánico en América, In Collard P. and Ubarri M. (coords.), *Encuentros, desencuentros, reencuentros: Flandes, Países Bajos y el mundo hispánico en los siglos XVI-XVII*, pp. 45-72.

A DH-Leavened Musicological Toolbox

Francesca Giannetti
francesca.giannetti@rutgers.edu
Rutgers University

Anna Kijas
kijas@bc.edu
Boston College

In 1996, Don Michael Randel wrote a chapter entitled "The Canons in the Musicology Toolbox," which examined "a common set of techniques that every dissertation and scholarly article employs" a type of "musicological interface" or "toolbox" that addresses the issues of theoretical and methodological consistency across the discipline, and, in theory, reduces the time and effort spent on producing the scholarly output (Randel, 1996:10). Fifteen years after Randel's chapter was published, Zoe Lang reexamined his original concept in a post entitled "Today's Musicological Toolbox," in which she asks that we imagine afresh this concept of the "musicological toolbox" (Lang, 2010). Her argument was that a customized toolbox of both specialized and general tools would be required for producing a diverse range of musicological scholarship. The training of the musicologist should thus include as many of the specialized and generalist tools as possible, with the aim of creating a common ground with scholars of other disciplines, allowing the musicologist to move beyond the boundaries of her academic discipline. In the past twenty years, skills once considered general and essential (music notation) have given ground to areas of specialization like popular music and feminist readings. As in other humanistic disciplines, musicologists consider postcolonial and identity theories, together with the more "native" topics of music analysis and Western "art" repertories.

If we were to reimagine the "musicological toolbox" yet again, how would we do so from a digital humanities and/or information science perspective? What tools should be added to a customizable "musicological toolbox" such that students and faculty become proficient in applying a set of techniques that may eventually become commonplace to all humanistic disciplines? Are there specific misapprehensions to be wary of when appropriating tools created for other academic fields? Can an incubator approach be tolerated to make room for learning and experimentation, without requiring formal, publishable results? Finally, how does one make room for collaboration in a discipline that is still largely driven by individual scholarly endeavor?

As librarians who study and practice both digital humanities and musicology, we use certain tools and methods that we propose be part of a DH musicological toolbox. These include free and open source tools for data capture, cleaning and formatting, geospatial, temporal and network analysis, as well as tools for data enhancement and encoding (i.e. metadata and text/music encoding).

We propose that concepts and methods core to digital humanities, as well information science, should be introduced into the musicological toolbox to expand students' abilities and understanding beyond the boundaries of musicology. These include:

- Digital curation and publishing
- Metadata standards and schemas
- Information architecture (i.e. relational databases, indexing)
- Close reading and content analysis (i.e. encoding (text/music), transcription, analyses)
- Graph theory
- Spatial and temporal analysis
- Exploratory data analysis for framing scholarly arguments

Training students and scholars in all of these concepts or tools may not be feasible, however as it is important for musicologists to be introduced to methods, such as historiography, paleography, or musical analysis, it is also important that they are introduced to the concepts and methods used in digital humanities work, which will allow them to push the boundaries of musicological research and build an understanding around developing scholarship or research projects in a digital mode.

Drawing from our own experiences, first as musicology students, and now as librarians, we will demonstrate how these concepts may be applied to musicological research using our current projects as case studies, *Documenting Teresa Carreño* (Kijas) and an analysis of librettist Felice Romani's *I due Figaro* (Giannetti). In addition, we will demonstrate some of the rewards and challenges of blending the information science, musicology and digital humanities perspectives.

In *Documenting Teresa Carreño*, a project focused on the performance career of Teresa Carreño (1853-1917), a Venezuelan pianist and composer, Kijas curates metadata and content in Omeka to document Carreño's key performances between 1862 and 1917. An understanding of concepts and methods related to metadata standards, musicological research, as well as, digital curation and publishing, are especially relevant to this project. Giannetti's project analyzes Felice Romani's *I due Figaro* alongside its French source play, *Les deux Figaro*. The concepts and methods applicable to her project include close reading, OCR, text transcriptions, network, data and text analysis. These two case studies demonstrate a data-enhanced view of the musicological toolkit in which findings drawn from traditional approaches (archival studies, bibliography, close reading) are challenged and supplemented by digital concepts and methods. Although not without its complement of additional labor and uncertainty, this DH-leavened musicological toolkit compensates by improving one's understanding of sources, enhancing digital literacies, and raising the chances of finding intra- or inter-departmental collaborators.

Bibliography

Lang, Z. (2010). Today's Musicological Toolbox. *Amusicology*. https://amusicology.wordpress.com/2010/02/05/guest-post-by-zoe-lang-today%E2%80%99s-musicological-toolbox/.

Randel, D. M. (1992). The Canons in the Musicological Toolbox. In Bergeron, Katherine and Philip V. Bohlman (Eds.), *Disciplining Music: Musicology and Its Canons*. University of Chicago Press.

The Garden: A 3D Adventure Puzzle Game Exploring Bosch's Garden of Earthly Delights

Elizabeth Goins
esggsh@rit.edu
MAGIC Center, Rochester Institute of Technology, United States of America

Christopher Egert
caeics@rit.edu
MAGIC Center, Rochester Institute of Technology, United States of America

Andrew Phelps
amp5315@rit.edu
MAGIC Center, Rochester Institute of Technology, United States of America

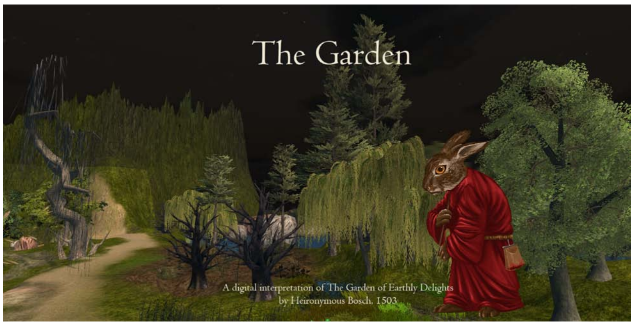
The design of serious, humanities-rich games is a challenging process full of difficult problems. Of particular interest to us is the development of humanities games that might appeal to the general public and compete with commercial games. *The Garden*, is a high level prototype that allows players to explore the world of Hieronymus

Bosch and his painting, *The Garden of Earthly Delights*. *The Garden* prototype focuses on concept, level design and game mechanics and was submitted for review in competitions such as IndieCade. Although the level of polish of the prototype was not high enough for inclusion in mainstream festivals, the concept and game elements were enthusiastically received by the jurors. Our development focus has now shifted to review of the educational content and the development of narrative and characters before the final phase of art asset creation and polish. We are looking for input and feedback from the digital humanities community to improve the educational content and gameplay. Visitors to the demonstration will be able to play the prototype, complete surveys and join in the discussion to help make this a better game for humanities outreach.

To play the game, players take on the role of Godefroy, a tormented denizen of Musician's Hell who is confined by fiery borders within a small patch of the underworld. As Godefroy, the player tries to escape hell by figuring out the secret of each NPC in order to possess them. This is the only way for the player to interact with different areas of hell and win the game.

Mechanics are used to connect the game play to the educational goals of the project. One example is that of engaging the player with history and material culture through the manipulation of identity. As the game location is Musician's Hell, each of the NPCs is drawn from a diverse cast of historic troubadour or trobairitz. General audiences have preconceptions about the Middle Ages and Early Renaissance that are limited, and sometimes incorrect, that they have learned from popular culture. The possession mechanic was developed specifically to engage players with these characters in order to deepen understanding of the culture: musicians could be females or Jews, for example, with rich histories and personal lives. Not only do the players engage in dialogue, but they also must possess each of the characters and take on their identity in order to progress the game. This gives an opportunity to see the world from that NPC's perspective.

The game environment is filled with themes derived from the painting itself. The player is immersed in a world where everything is watched and counted. The object in this interpretation was not to recreate the painting but to interpret it in a way that would resonate with the contemporary players. Players encounter an underworld different from those commonly depicted so that they may connect with Bosch's themes: For example, the importance of being able to see past the superficial, material surface to the spiritual truth of worldly choices and actions. In *The Garden* this is expressed by the environment. Here, Hell is beautiful and invites exploration but poisonous secrets are buried beneath the façade. The player must search for a way out by jumping from soul to soul while avoiding the relentless demon horde run by the bureaucrats of Hell. Along the way, to face the awful overlord Mallebisse, they uncover a story of lust, jealousy and betrayal. Overall, the game puts them in the position of questioning who they really are and what the ultimate objective is in a way, we hope, that parallels Bosch's understanding of choice and salvation.


The Garden
A digital interpretation of The Garden of Earthly Delights
by Heironymous Bosch, 1503

Authorial {X}: A Research and Teaching Platform for Literary Geography

Karl Grossner
karlg@stanford.edu
Stanford University, United States of America

Kenneth Ligda
kenligda@stanford.edu
Stanford University, United States of America

The Authorial {X} project introduces novel means for compiling and mapping references to place within literary works, then exploring and analyzing them from literary, geographical, and biographical perspectives. Its first exemplar locale is London, significantly extending a 2011 project directed by Professor Martin Evans (2011), which mapped the residences of 47 London authors living between the 14th and 20th century. The revamped *Authorial London* project (https://authoriallondon.stanford.edu) delves deeply into hundreds of written works by those same authors, with interactive features permitting the examination of over 1600 place-inflected passages, faceted on dimensions of genre, form, literary community, social standing, and neighborhood.

The *Authorial London* web application has been developed as a re-usable platform, hence, "Authorial {X}.“ Our intent is that researchers and college instructors may readily instantiate a similar site for any place of interest, and engage either a class or research community of interest in gathering data for it: authors, texts, references to places within texts, and the geometry necessary for mapping them. The software code and documentation required for standing up an empty, configurable Authorial {X} instance will be freely available on GitHub.

Scale

The Authorial {X} platform operates at multiple scales of analysis, viewed from both geographic and literary perspectives.

Geographic scale

Cities are a distinctive kind of place in many respects, sharing the quality of geographic variability of lived experience with areas of other scales. Neighborhoods within cities are distinctive places themselves: Belgravia and Southwark are both in London, but are not similar—either in economic and social scientific terms, or in the literary imagination that is this project's focus. The same can be said of provinces within countries and districts within provinces. Authorial {X} provides one means for discovering within literature how places within places of any scale differ, and whether and how they have changed over time.

Analytic scale

The platform is designed to permit the close reading of particular works, at the scale of passages, and also facilitates the aggregation of those passages in several ways: across *authors*, by any of 17 genres (e.g. comedy, novel, bildungsroman) and 3 forms (prose, drama, and poetry); across *time*, by one of 13 literary communities (e.g. Romantic, Victorian, Modernist); and across *space* (by several dozen neighborhoods). These groupings are visualized *spatially*, as interactive dots and lines on a map, *temporally* in simple histograms; and *conceptually*, in word clouds of salient terms as measured by the TF-IDF statistic computed on the project corpus.

Data

Textual

The *Authorial London* corpus has been developed by manual and machine-assisted means. Initially, over 600 passages were hand-selected from approximately 80 works under copyright. The place references tagged in those passages were then used to search an indexed corpus of 690 texts in the public domain acquired from Project Gutenberg. The result narrowed the search space dramatically (from 690 to 220 texts), allowing the still arduous manual selection of several hundred more passages in a timely fashion. During the second round of manual selection numerous additional place references were identified, and the process repeated. Passage length varies significantly, from a sentence or two to several hundred words.

In addition to passages from authored works, the application presents georeferenced biographical essays for each of the 47 authors.

Geographic

All references to places—whether in featured works or biographical essays—are "place-reference" records, each linked to a spatial location. Their scale ranges from a single tavern to the entire city. A place may have multiple names, but link to the same geometry; for example "The Thames" and "the silver Flood" are from different works, referring to the same place and physical location. Cartographic representations are not limited to single points; many streets are displayed as lines. A modern basemap is supplemented by three geo-rectified ("warped") historical maps.

Graphical

A number of photographs related to locations within the biographical essays are also made available in the interface.

Platform details

The Authorial {X} platform back-end is built with the Ruby-on-Rails framework on a PostgreSQL database. Its custom front-end code leverages several JavaScript software libraries, including Leaflet for mapping and D3 for visualization. The Rails technology has enabled the efficient development of an administrative interface to the data, wherein a set of simple web-based forms for populating and editing the database are available to authenticated users. In this way, once an instance of Authorial {X} is initialized, non-technical users can develop data that will appear dynamically in the interface. The project is open-source, and its developers expect that improvements will be made by digital humanities developers over time.

Bibliography

Evans, M. and Meeks, E. (2011). *A Guide to Authorial London.* http://authorial-london.stanford.edu (accessed 25 February 2016).

Documenting the pain: Sharing Second World War survivors' stories to help meaning making and lessons learning through curating trans-European digital narrative trajectories

Siegfried Handschuh
siegfried.handschuh@uni-passau.de
Universität Passau, Germany

Simon Donig
simon.donig@gmx.com
Universität Passau, Germany

Adamantios Koumpis
adamantios.koumpis@gmail.com
Universität Passau, Germany

Hanna Diamond
DiamondH@cardiff.ac.uk
Cardiff University, U.K.

"The world is the totality of facts, not of things" but facts do not speak for themselves; they have no voice. The life-world is the totality of human experiences. When people die, those experiences are lost, unless they are recorded in some way, becoming part of individual or collective memories. This includes Second World War survivors' experiences of the role of the war in their lives, over their subjective time, and the physical and social spaces that they have traversed. With our research, we aspire to explore novel ways to capture, preserve, curate, organise and communicate this set of experiences, stories, narratives, so that they can constitute a shared resource that people can augment, and that individuals as well as institutions can delve into, to find inspiration for new ways of conceptualising acceptance, tolerance and understanding, and how these new ways can be reflected into every-day practices and policies, and foundations for visions of our future.

The Second World War was one of the major transformative events of the 20th century. Millions of lives were lost crimes were committed, physical capital was destroyed, populations were displaced, families experienced extended periods of separation, and many Europeans, including young children were exposed to the horrors of War. Unlike many earlier wars when casualties were mainly confined to the battlefield, during this conflict civilians were also directly affected by warfare. About half of the European casualties of the war were civilians (including women and children). At the end of the war, millions of people were homeless, the European economy had collapsed, and much of the European industrial infrastructure had been destroyed. Reconstruction in the aftermath of the Second World War brought significant change to political, economic and social systems across most European countries and was the touchstone for the realization of the European project.

The trauma of these Second World War experiences and the upheaval it brought to ordinary people's lives continues to have a compelling resonance in contemporary European societies. Formal commemorative practices on local, national and international platforms draw attention to the bravery of combatants and pay tribute to those who lost their lives, during this and other conflicts, as demonstrated by the recent commemorations of the centenary of the First World War and the 70th anniversary of the Second World War. As the Second World War survivor generation disappears, their lived experience is also being lost. However, family story telling means that their personal stories continue to circulate. Individuals who have passed on their personal war stories to their descendants, along with these descendants, comprise contemporary European society and shape 'European identity'. From a psychological point of view, citizens experience all these war-related political and economic changes on a personal level and they develop their own personal stories which may differ from official political discourses. Since past stories transform our present image of the war, and past stories can save both us and our descendants from committing the same 'mistakes' in the future, it is important to uncover, save and preserve individual's personal and family war stories, and officially record them as a part of European cultural heritage. Our research seeks to develop a story telling platform and research framework which will allow for the collection of these stories. This will process material and narratives from three source types:

1. Resources provided by Second World War survivors which were originally collected at source / by the source (e.g. diaries, letters),
2. Resources provided by Second World War survivors that were later created in the form of memoirs or documentaries,
3. Resources created by third parties (family, children, others) and which may be of a variety of types like fiction / non-fiction works, scientific publications in the form of e.g. research dissertations, journalistic newspaper articles, posts in social media or blogs, etc.

They will be complemented by contributions from multimodal sources such as open data collections, closed archives, collections of pictures (drawings and photographs), video footage and sounds, etc.

We aspire to offer a unique access point for producing and annotating stories on the basis of impromptu / ad hoc ontologies that will be leveraged to construct trajectories of narratives over time. Besides a time-based order of the narratives, ontologies could relate key concepts, key emotions, and key arguments, in single narratives or over several narratives by the same or by different individuals. Part of the work can already start with narratives on existing community platforms exploiting the www.fleeinghitler.org infrastructure.

The Ibn Darraj Project: SpatioTemporal Reasoning Engine Based on Evidence Combination

Mohammad Hossein Haqiqat khah
mh.haqiqatkhah@ut.ac.ir
Machine Learning and Computational Modeling Lab, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran

Babak Nadjar Araabi
araabi@ut.ac.ir
Machine Learning and Computational Modeling Lab, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran

Introduction

In historical statements, we face narrations that are vague in sense of persons, places, times, and events. This means that we have uncertainties and non-specificities about the persons and the events, and the time and place the event has taken place. These uncertainties are mostly caused by ignorance about the characters and events assignments, and definition of the boundaries narrowing down the time and space frames. Moreover, we have narrations that usually do not uphold each other, and have different validities or certainties. Putting these contradictory narrations together to judge about the most definite event needs a great effort, and is a hard task for unaided human mind.

Although its importance and utility, there are not any noteworthy research on spatio-temporal reasoning to combine evidences, or applying evidence combination methods (such as Dempster-Shafer Theory of Evidence) in spatio-temporal reasoning. Hence, this paper delivers a novel tool to compensate this deficiency.

Methodology

To reach this goal, we divide the problem to three main domains, and simplify the question to minimize the efforts and implementation cost of each step.

Extracting and Structuring Evidences

In this step, we model narrations as narration trees, such as *'A' says he has heard 'B' talking about [Event]*. Each of the narrators has a validity coefficient indicating the reliability of the narration made by the person. Then using the SRL (Semantic Role Labeling) methods (Gildea and Jurafsky, 2000), we structure the narrations to a standard form Jurafsky and Martin (2009). Hence, we have a bunch of narrations, and each of the narrations has a narration chain and a standard, structured event.

Finding Similar Events

In this step, using Natural Language Processing tools, we form [multiresolution] verb clusters that help us define the same events that have meanings in common. For example, killing, shooting down, hanging, and choking can be considered as different ways to end one's life, and we may put them in one cluster. These clusters are extracted from WordNet (Hirst and St-Onge, 1998), and define the degree to which we may merge events based on their verbs (Meng et al., 2013).

Evidence Combination

The main role of the reasoning engine is to combine different evidences to calculate the probability of a hypothesis. To do so, one can use the very familiar Bayesian inference method to incorporate different evidences using the Bayes theorem. However, this modeling cannot deal with non- specificities. For instance, if we have four persons A, B, C, and D as candidates of an assassination, we may assign probabilities to Assassins = {A,B,C,D} or other probabilities to subsets of the Assassins set. However, if we do not have crisp evidences defining the probabilities of each member we are unable to do inference where the subsets overlap. As a result, we cannot use the Bayes rule to combine evidences of Pr(A or B or C) and Pr(B or C or D) if we do not know Pr(B or C)or assume some other hypothesis for it.

On the other hand, there is a great advantage in using the Dempster-Shafer Theory of Evidence (DST) compared to the classical probabilistic reasoning based on the Bayesian Theory (BT). By DST, it's possible to take non-specificity into account as well as randomness, which was not possible in BT. In many problems, such as reasoning based on non-specific statements (which are modeled here as narrations), accumulated non-specificity (vagueness) may reveal more specific details. This is the main reason to use DST instead of BT.

It means that when each narration determines a non-specific time and place for an event, or these space-time pieces overlap each other, we are able to define precisely the most and least probabilities that the event has taken place in a specific portions of space-time. These probabilities are usually mentioned as lower and upper bounds of probability, and may be interpreted as Belief and Plausibility respectively (Zadeh, 1986).

Tools and Methods

The core element of the system is a reasoning engine that combines different pieces of evidence. The output will be total expected probabilities and the upper and lower bound of probabilities for different pieces of space-time. This task can be accomplished by efficiently succeeding in the following:

- Modelling space and time effectively, and
- Using the evidence combination methods to deliver the reasoning engine, mainly using Dempster-Shafer Theorem.

We used the CRMgeo model (Doerr and Hiebel 2013) that is standardized in ISO 21127:2014 as the space-time model. We have incorporated GIS tools (as done in Hirschfield and Bowers, 2001; and Fuhrmann et al., 2013) and standard ISO 8601:2004 to model space and time respectively. However, due to the rich and flexible ontology design of the CRMgeo model (Doerr and Hiebel, 2013) which is standardized in ISO 21127:2014, we switched to it as our main space-time model. We also benefited from the CRMinf argumentation model (Doerr 2015) to model the narrations.

At last, we implemented different evidence combination rules in R language to combine the structured narrations together, resulting in and reduced ambiguity and vagueness (Kohlas and Monney n.d., Barnett 1991).

An Example of a Toy Problem

The following figures are the output results of a reasoning over a toy problem of an airplane crash in different boroughs of London, and is plotted for each of the years in the 1930-1935 timeframe. Each plot is for a specific time (e.g. X1930 is for the year 1930). The values of the average probability of the crash in each borough are represented in decibels to better visualize the slight changes of probability in similar regions.

The problem is to combine the following five narrations to gether. The numbers in the parentheses are the validities and/or confidence of the statements.

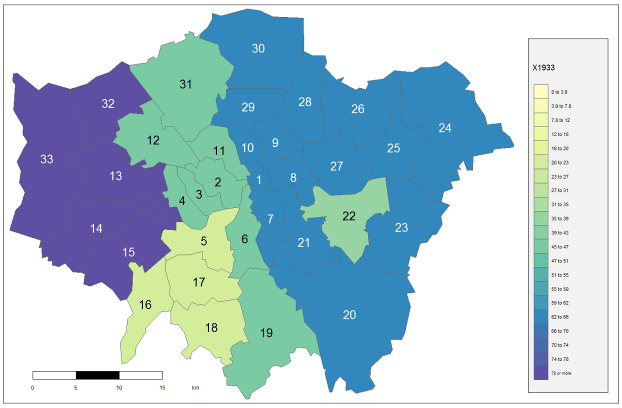
1. I think (70%) it was between 1930-33 that a plane crashed in the south east of London

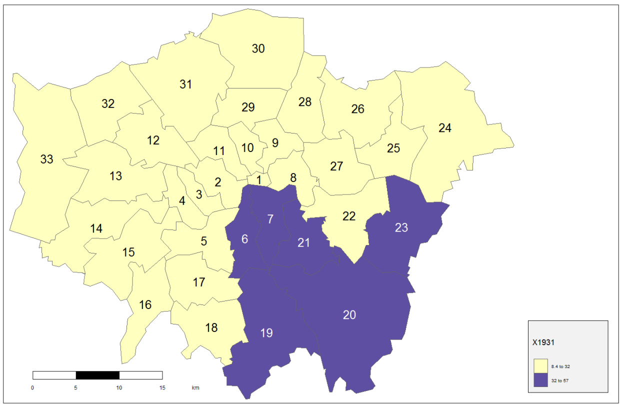
2. If I'm correct (90%) I heard my brother that he's somehow sure (80%) that it was between 1932-35 that the Air Union cargo plane hit the northern bank of the Thames river.

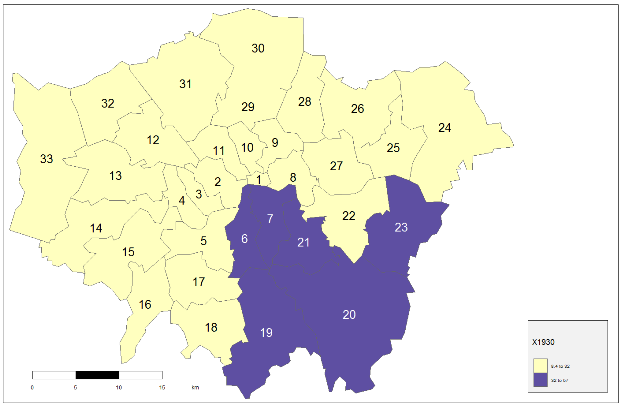
3. I cannot remember it clearly (60%) but in 1933 or 34 an airplane of a post company hit the eastern London.

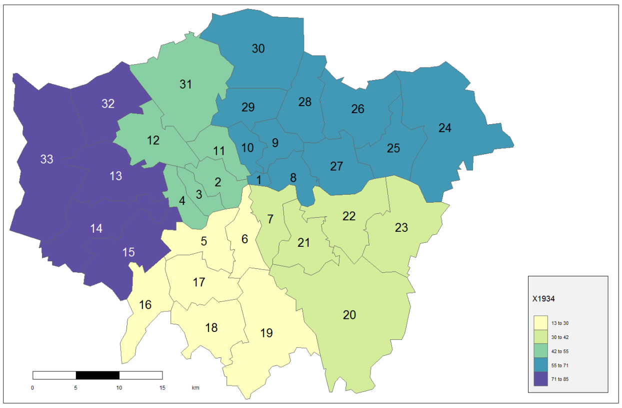
4. It's hard to remember (60%), but my father once told me (90%) that he witnessed an airplane crash in the center of London, Old London.

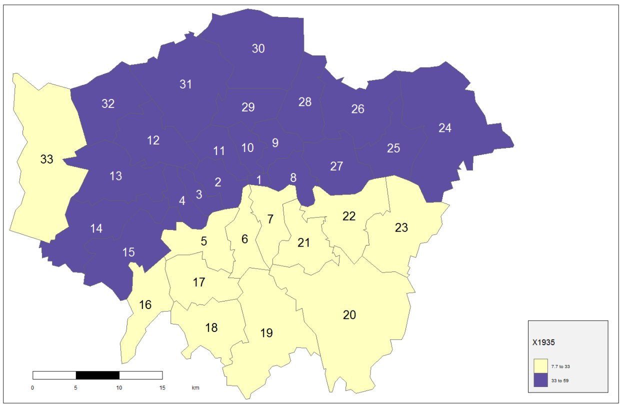
5. If I'm remembering correctly (80%), an airplane crashed the west London between the 1932-34.











About the Project Name

Nuh (Noah) ibn Darraj al-Nakhaʿı¯ (d. 182 AH/ AD 798), was the Shıʿite judge of Kufa, and later in his life, the grand judge of the eastern half of Baghdad. *Nuh*'s older brother, *Jamil ibn Darraj al-Nakhaʿı¯* was a prominent Shıʿite jurist in the latter part of the second Islamic century (AD 760s - 810s) (Modarressi, 2003)

Bibliography

Barnett, J. A. (1991). Calculating Dempster-Shafer plausibility. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**(6): 599–602.

Doerr, M. (2015). *CRMinf: the Argumentation Model.*

Doerr, M. and Hiebel, G. (2013). CRMgeo : Linking the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement, pp. 1–40.

Fuhrmann, S., Huynh, N. T. and Scholz, R. (2013). *Crime Modeling and Mapping Using Geospatial Technologies*, Available at: http://link.springer.com/10.1007/978-94-007-4997-9.

Gildea, D. and Jurafsky, D. (2000). Automatic labeling of semantic roles. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*, (1972), pp.12–520. Available at: http://portal.acm.org/citation.cfm?doid=1075218.1075283.

Hirschfield, A. and Bowers, K., (2001). *Mapping and analysing crime data*, London: Taylor and Francis, 2001. Available at: http://discovery.ucl.ac.uk/1329199/.

Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet - An Electronic Lexical Database*, pp.305–32. Available at: http://mitpress.mit.edu/books/wordnet.

Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*. Pearson.

Kohlas, J. and Monney, P. A. (2013). *A mathematical theory of hints: An approach to the Dempster-Shafer theory of evidence*, **425**, Springer Science and Business Media.

Meng, L., Huang, R. and Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology*, **6**(1): 1–12.

Modarressi, H. (2003). *Tradition and Survival: A Bibliographical Survey of Early Shi'ite Literature*, **1**.

Zadeh, L. A. (1986). Simple View of the Dempster-Shafer Theory of Evidence and Its Implication for the Rule of Combination. *AI Magazine*, **7**(2): 85–90. Available at: http://www.aaai.org/ojs/index.php/aimagazine/article/view/542.

The Kuzushiji Project: Developing a Mobile Learning Application for Reading Early Modern Japanese Books

Yuta Hashimoto
yhashimoto1984@gmail.com
Kyoto University, Japan

Yoichi Iikura
iikura@let.osaka-u.ac.jp
Osaka University, Japan

Yukio Hisada
fmptsr3431@gmail.com
Osaka University, Japan

SungKook Kang
izaya6013@yahoo.co.jp
Osaka University, Japan

Tomoyo Arisawa
t.arisawa212@gmail.com
Osaka University, Japan

Akihiro Okajima
aki@o.email.ne.jp
Osaka University, Japan

Tsutomu Yada
mtsutomu@tiara.ocn.ne.jp
Osaka University, Japan

Rintaro Goyama
goyama@let.osaka-u.ac.jp
Osaka University, Japan

Daniel K-B
wappuccino@gmail.com
Osaka University, Japan

Introduction

It frequently happens in the modern history that a certain cultural skill that used to be shared in a community or society is lost in the process of modernization. Over the last century, Japanese people have lost the basic skills to read *kuzushiji*, classical calligraphic renderings of Japanese characters (see Fig. 1). Since Japanese society ceased to use *kuzushiji* for publication after the Meiji Period (1868-), most Japanese today except trained experts are unable to read books printed with *kuzushiji* only 150 years ago.

However, from 2008 a group of scholars of Japanese literature have started calling attention to the need for providing younger generation with the tools to access their own past (Nakano, 2011 and Moretti, 2014). The *kuzushiji* project, which started in 2015, is an attempt to build a mobile learning application that provides resources and trainings for reading *kuzushiji*. We call the app the KuLA (*kuzushiji* learning application). It is a public humanities project, as well as an interdisciplinary project of Japanese Literature and Digital Humanities scholars. In this paper, we will briefly describe the features, workflows, and implementation of the KuLA.

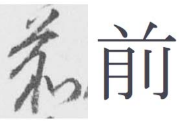
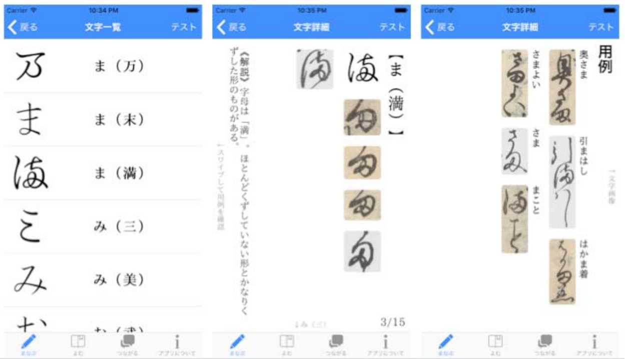


Fig. A comparison of a *kuzushiji* and modern Japanese type font. Both represents the same character 前, "front" in English



ま（万）
ま（末）
ま（満）
み（三）
み（美）

【ま（満）】

《解説》字母は「満」。ほとんどくずしていない形とかなりくずした形のものがある。

3/15

用例

奥さま

引まはし

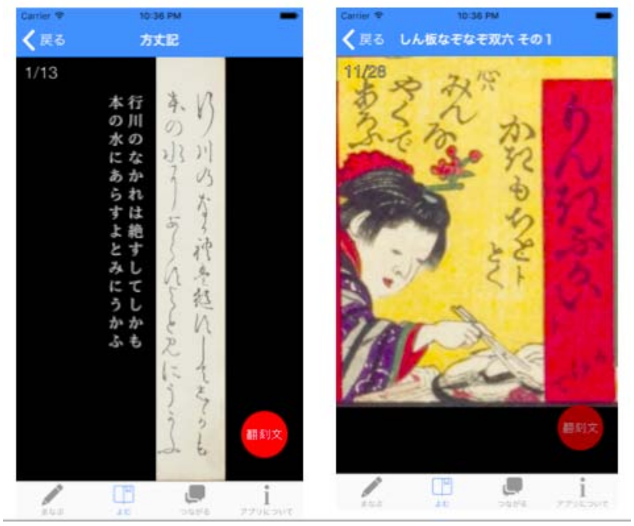
はかま着

さまよい

さま

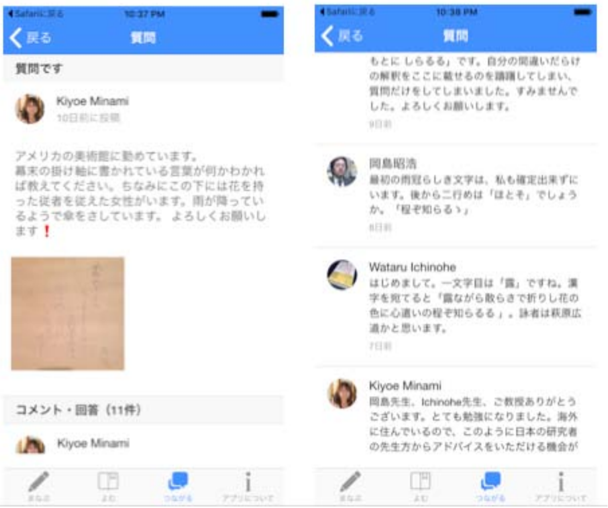
まこと

Fig. The character module



行川のなかれは絶すしてしかも本の水にあらすよとみにうかふ

Fig. The reading module



質問です

Kiyoe Minami
10日前に投稿

アメリカの美術館に勤めています。
幕末の掛け軸に書かれている言葉が何かわかれば教えてください。ちなみにこの下には花を持った従者を従えた女性がいます。雨が降っているようで傘をさしています。よろしくお願いします❗

コメント・回答（11件）

Kiyoe Minami

もとに しらるる」です。自分の間違いだらけの解釈をここに載せるのを躊躇してしまい、質問だけをしてしまいました。すみませんでした。よろしくお願いします。
9日前

岡島昭浩
最初の冊冠らしき文字は、私も確定出来ずにいます。後から二行めは「ほとそ」でしょうか。「程ぞ知らるゝ」
8日前

Wataru Ichinohe
はじめまして。一文字目は「露」ですね。漢字を宛てると「露ながら散らさで折りし花の色に心遣いの程ぞ知らるる」。詠者は萩原広道かと思います。
7日前

Kiyoe Minami
岡島先生、Ichinohe先生、ご教授ありがとうございます。とても勉強になりました。海外に住んでいるので、このように日本の研究者の先生方からアドバイスをいただける機会が

Fig. The community module

Features of KuLA

Learning *kuzushiji* is essentially similar with learning a foreign language. What you need for learning it are as followings: 1) the basic knowledge about *kuzushiji*, especially of character shapes, 2) decent amount of exercises of reading actual texts, and 3) good mentors and fellows who will teach and motivate you.

For the reasons above the KuLA consists of the following three modules:

- **Characters module**, where the user will learn the basic knowledge about *kuzushiji*, especially about character shapes. The user can browse the list of *kuzushiji* characters and jump to the detail page of each character (see Fig. 2).
- **Reading module**, where the user will read actual texts written with *kuzushiji* for exercises. The user can also check their transcribed texts (see Fig. 3).
- **Communication module,** where the user will communicate with others via the network. This module will, for instance, enable the user to ask others how to read a specific *kuzushiji* characters by sharing photos taken by the user (See Fig. 4).

Workflow and Implementation

In order to create the teaching materials bundled with the KuLA we needed a lot of actual images of kuzushiji. For collecting them efficiently we developed a Chrome extension which enables to capture arbitrary image regions from the digital collection of pre-modern books provided by the National Institute of Japanese Literature. The images captured by the extension will be automatically uploaded to an web app built with Ruby on Rails. In this way we have gathered so far about 3,000 images of kuzushiji characters.

The mobile app was built with Ionic, a HTML5 mobile framework based on Apache Cordova and AngularJS frameworks. The use of HTML5 technology makes it possible to generate the distributions both for iOS and Android from a single source code. As the backend of the mobile clients we used parse.com, which provides basic server-side features such as user authentication and data storage.

Conclusion and Future Directions

KuLA was released both on Google Play and on App Store for free on 18 Feburary 2016. It has been downloaded more than 5,000 times in two weeks after the launch. The average review scores are 4.5/5.0 in AppStore (total 15 reviews), and 4.9/5.0 in Google Play (total 29 reviews). From these numbers we may say that our design strategy for building KuLA was successful.

We believe that it is a duty of humanities scholars to build the tools to access the past knowledge for further generations. And what makes Digital Humanities special in this regard among other humanities discipline is that it can directly provide those tools with the public with the help of digital media such as mobile devices.

Bibliography

Nakano, M. (2011). *Wahon No Susume*. Iwanami Shoten.

Moretti, L. (2014). *Reading hentaigana and kuzushiji Manual.* http://wakanedo.com/wp/wp-content/uploads/2014/07/Edo-no-kakikotoba-2014-hentaigana-kuzushiji-Manual-1.pdf

Notes

[1] Our project blog: https://plus.google.com/104467959383842469455/posts.

[2] https://www.nijl.ac.jp/

[3] You can see the list of kuzushiji images we have collected in the following link: https://youreicollector.herokuapp.com/characters/

[4] http://ionicframework.com/

Blacks In American Medicine Archive: Exploring Forgotten Stories

Evan Higgins
elh@mit.edu
MIT HyperStudio, United States of America

Kurt Fendt
fendt@mit.edu
MIT HyperStudio, United States of America

Josh Cowls
cowls@mit.edu
MIT HyperStudio, United States of America

Andy Stuhl
akstuhl@mit.edu
MIT HyperStudio, United States of America

Archival work has for centuries privileged the aggregation over the dissemination of rare and important content. And while this has been necessitated largely in order to preserve the objects within, it has particularly been harmful to marginalized communities whose narratives lie outside of mainstream consciousness. Thankfully, the affordances of digital humanities has opened up new avenues for these elusive materials, and the expansive histories they hold within, to reach a multitude of audiences.

The newest project at HyperStudio, MIT's Digital Humanities Center, the Blacks in American Medicine (BAM) archive innovates on this unique potential offered by digital media to promote and display never-before-seen materials. The BAM project features over 23,000 biographical records of African American physicians from 1860-1980 and countless primary documents associated with these practitioners and the African-American medical community at large. From numerous pieces of personal correspondence, such as a letter to the AMA pleading with them to hold the annual convention in a non-segregated locale, to unique biographical data that charts the ebb and flow of African Americans in medicine, much of the content within the BAM archive has never been available to a wide audience, and all of it has never been digitalized in a central location. Our archive incorporates both a focused study into the history of specific physicians and a broader analysis of the trends within the African American medical community to unearth untold chapters in the vast history of the black experience in America.

While still in the initial stages of this project, we are working on a number of intersectional methods to display this wealth of content. As with most of HyperStudio's archival projects, we are making sure that the content is discoverable by both scholars as well as more casual audiences. This begins by making sure that the content is encoded using established metadata standards such as Dublin Core, allowing to connect our high resolution primary materials and biographical records to other relevant archives. Additionally, within our site itself, we plan to integrate our Repertoire faceted browser, which allows for both a targeted search usingspecific criteria and the ability to explore interconnected documents that interest the user. Additionally, this project will feature our Chronos Timeline, which dynamically uses events and occurrences to present historical data. Outside of displaying content algorithmically and chronologically, we plan on incorporating geographic, visual and biographical tools to help novices and experts alike delve into this content in order to discover new stories and explore existing narratives. By allowing users multiple entry points, these cross-sectional methods of content interrogation will further highlight our project's unique ability to test traditional views of the African American experience, which focus on a few key moments that affected the larger populace than the long history of the people themselves.

Our job as custodians of this trove of content is to make sure that it is not only widely accessible but, more importantly, that it is intuitive and useful to our audience. A poster session at DH2016 will allow us to gather feedback from thought leaders in the field in order to facilitate the evolution of our product. This is a crucial step in making sure that our project has the impact, reach and power that it deserves.

At HyperStudio, the Massachusetts Institute of Technology's center for digital humanities, we use new media to discover and explore forgotten stories. By both focusing on specific people, places and details and zooming out to view trends and patterns, our methods and tools allow us to question traditional narratives and investigate emerging ones. This project, a collaboration with Pulitzer-Prize-finalist author and African-American science historian, Kenneth Manning, demonstrates the power of new/innovative archival approaches to discover and promote new content. This archive is a chance to articulate stories that have remained untold and question narratives that have remained unchallenged, and the Digital Humanities 2016 conference is the perfect place to deepen this discussion.

Abbreviations In Manuscripts – Systematization And Crowdsourcing By Ad Fontes

Tobias Hodel
tobias.hodel@uzh.ch
Universität Zürich, Switzerland

In an unprecedented and very successful crowd sourcing project, the most important resource for abbreviations in Latin and Italian has been digitised and is now freely available. The e-learning project Ad fontes managed to accumulate and systematize all 14'357 abbreviations contained in the most renowned collection, the *Dizionario di abbreviature* by Adriano Cappelli (Cappelli, 1928). The digitized and systematized abbreviations offer new ways to access handwritten texts. Besides enhanced search methods (with wildcards, uncertainty, abbreviation placement etc.) it will soon be possible to add new abbreviations and rectify entries that were handled incorrectly by Cappelli.

The poster presents the crowd-sourced digitisation of the printed Lexicon and considers specific problems and lessons learned while dealing with the crowd; it gives insights into new possibilities regarding the research of abbreviations, discusses possible ways to deal with abbreviations, and at the same time raises questions concerning requirements and nice-to-haves for an application that hopes to become an ever-growing resource to abbreviations found in historical sources.

Following Cappelli's model, each abbreviation is presented as a „facsimile" of a hand-written abbreviation followed by a transliteration of the letters present, the placement of the abbreviation symbol(s) in a grid, if applicable the category of context (e.g. legal or medical) as well as the period the manuscript stems from.

All this information could be identified by users without further knowledge of paleography or Latin and thus offered ways to allow non-experts to take part in the crowdsourcing process. In addition to Cappelli's system, we had the participants place the abbreviation symbols within a 3x3 grid. This allows the introduction of a new search parameter refining the search according to the positioning of abbreviation marks. Unlike the data indicated in Cappelli (for which only a very limited number of mistakes and typos were entered), the placement of abbreviation marks was highly problematic and needed to be corrected in most cases by expert validation. In a next phase we would not include this part in the regular crowd sourcing process, but instead have it done separately.

Within 23 days, all of the abbreviations contained in Cappelli were digitised by our crowd sourcing participants. Mobilising such a crowd was made possible by a highly connected academe and archivists (via mailing lists such as digital medievalist as well as social media). Apart from the option of remote online participation, we planned a crowd sourcing event at the University of Zurich that was supplemented by smaller events at universities in Oxford and Berlin.

The information gained allows – based on very provisional data – conclusions concerning the use of abbreviations in Latin manuscripts. Of the 9'000 most common words in Latin (according to the SLU corpus, Pavur, 2009) 1094 (12.2 %, roughly 13% if we subtract words consisting of 3 letters or less which are usually not abbreviated) could potentially be abbreviated, only taking basic forms of words into account (no flections or conjugations). Compared to the Vulgate, the data shows that of the 38'138 words occurring in the Bible, 1083 exist in abbreviated form. Therefore the data demonstrates what could be expected (Traube, 1907): The abbreviations were not especially or solely conceived for the use in handwritten Bibles but for a variety of texts. A digital Cappelli is thus able to show in a quantitative way what specialists already suspected.

The abbreviations are being offered through the platform Ad fontes (www.adfontes.uzh.ch, Kränzle and Ritter, 2004) as well as the web app App fontes (t.uzh.ch/adf). A batch download of all data including the images is possible.

The digitization and systematization of the abbreviations according to Cappelli opens possibilities not included in the printed version: Unidentifiable letters and/or letters that are not part of the roman alphabet do not need to be known by the user; instead, they can use wildcards in order to get satisfactory search results. Generally, uncertainties won't prevent a successful search, they will only increase the number of results.

Currently, a feedback loop (concerning emendations of the Cappelli) as well as the possibility to add further abbreviations are being developed. Especially the second part will once more use the power of a specialized crowd, at the same time assembling a resource for the use of everyone working with manuscripts.

By crowdsourcing the digitisation of the Cappelli by a heterogeneous group of people, we proved that there is an interest in and a need to deal with abbreviations (Pluta, 2016). With its new resource, Ad fontes will facilitate how abbreviations in handwritten documents are dealt with in the 21st century.

Bibliography

Cappelli, A. (1928). *Lexicon abbreviaturarum: Wörterbuch lateinischer und italienischer Abkürzungen wie sie in Urkunden und Handschriften besonders des Mittelalters gebräuchlich sind.* Leipzig: JJ Weber.

Kränzle, A. and Ritter, G. (2004). Ad fontes. Zu Konzept, *Realisierung und Nutzung eines E-Learning-Angebots*. Zürich (http://opac.nebis.ch/ediss/20050043.pdf, accessed 4 March 2016).

Pavur, C. (2009). *Latin Vocabulary: High-Frequency Latin Word-Forms* [web site]. http://www.slu.edu/colleges/AS/languages/classical/latin/tchmat/grammar/vocabulary/hif-ed2.html (accessed 4 March 2016).

Pluta, O. (2016). Abbreviationes^{TM} Online – Medieval Abbreviations on the Web [web site]. (http://www.ruhr-uni-bochum.de/philosophy/projects/abbreviationes/index.html, accessed 4 March 2016).

Traube, L. (1907). *Nomina sacra. Versuch einer Geschichte der christlichen Kürzung*. München: CH Beck'sche Verlagsbuchhandlung (Quellen und Untersuchungen zur lateinischen Philologie des Mittelalters 2).

Annotating and Georeferencing of Digitized Early Maps

Winfried Höhn
winfried.hoehn@uni.lu
University of Luxembourg, Luxembourg

Christoph Schommer
christoph.schommer@uni.lu
University of Luxembourg, Luxembourg

Original early maps are usually only accessible for a small group of researchers and librarians because they are very old and sensitive, and could be easily destroyed. However, they are a valuable knowledge source for historical research, because they are also political and cultural evidences of its time. In the age of Digital Humanities, online access and information search in digitized historical documents and early maps allows people from all over the world to work with such artefacts of cultural heritage. However, the digitization solely generates images of the artefacts without any access to the semantics of the documents.

For most digital libraries of early maps (e.g. http://www.oldmapsonline.org/) the available metadata include only information about the map, e.g. author, title, size, creation date. Unfortunately, there is only little information about the data contained in the map. Tools for information retrieval in digitized early maps need to support users in typical queries like for instance:

- Development of places over time
- Toponym changes in the course of the history
- Position of the place from an early map on a modern map, if the place still exists.

Place markers and their text labels contain this information, for instance the place type town, village, village with church or a factory or mill. This makes the annotation and georeferencing of place markers a crucial task. This task is at the same time very challenging due to the nature of the manually created early maps, which contain a high variance in the used symbols. This gets even more complicated by the fact that a single map can easily contain many thousands of place markers. Therefore, proper tool support and automation of the annotation and georeferencing are of interest.

The Referencing and Annotation Tool (RAT) supports annotators in the task of identifying place markers in a digitized early map and helps to create a link to a modern map with minimum effort. Because many different symbols are used as place markers and they differ across maps, the user needs to select a template for each type of place marker and to manually annotate a small subset of the map. Based on this, the templates get adjusted; all parameters for the template matching are calculated to automatically preselect place markers with high confidence and assign them the most likely of the possible types. These annotations can also be added by hand and the automatically generated annotations can be corrected.

RAT facilitates georeferencing by suggesting the most likely modern places based on an estimated mapping between the coordinates of the pixel- and geocoordinates of the already georeferenced place markers. The number of suggestions can be further restricted through a phonetic search to places with names sounding similar to the name given on the map. This allows for identification of the modern place name using the historic name if the spelling has changed but the names still sound similar.

Currently, RAT is in a prototype stage which we tested with a range of 16th to 18th century maps. For instance, one of the maps in the test set contained 3809 place markers. An area with 47 place markers was manually annotated. Based on this initial annotation 3399 place markers were identified and 3339 of them were correct matches.

In other words, 98.2% of the identified markers were correct and 87.7% of the existing place markers were semi-automatically identified. A detailed description of the implementation and the functionality of an earlier version of RAT can be found in (Höhn et al., 2013) but the latest version uses a new template matching algorithm specifically optimized for maps where text, rivers and other structures are frequently next to the place markers and can disturb standard template matching algorithms.

We plan to reduce the manual work needed in all areas by identifying similar maps of the same region and time and exploiting their similarities to provide better suggestions for existing places and their georeferencing on new maps.

Bibliography

Höhn, W., Schmidt, H.G. and Schöneberg, H. (2013). Semiautomatic Recognition and Georeferencing of Places in Early Maps, *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, Indianapolis*, pp. 335-38.

The »Hidden Kosmos« of scientific instruments in Alexander von Humboldt's »Vorlesungen über physikalische Geographie« – or: What on earth is an »Ariometer«?

Marius Hug
marius.hug@culture.hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Christian Thomas
christian.thomas.1@staff.hu-berlin.de
Humboldt-Universität zu Berlin, Germany; Berlin-Brandenburgische Akademie der Wissenschaften

Abstract

In 1827/28 Alexander von Humboldt presented the scientific knowledge of his time in lectures which became famous in their own right as a milestone in the popularization of science as well as due to their apparent connection to Humboldt's most important work "Kosmos. Entwurf einer physischen Weltbeschreibung" (1845–62).

The project "Hidden Kosmos—Reconstructing Alexander von Humboldt's 'Kosmos-Lectures'"[1] is located at Humboldt-Universität zu Berlin. Within two years (ending in August, 2016) the project digitizes and virtually brings together eleven attendee's lecture notes (all that are known at the moment) of the "Vorlesungen über physikalische Geographie", the so-called "Kosmos-Lectures", coming from libraries, archives as well as private collections in Germany, Poland (Kraków) and Turkey.

The TEI-encoded transcripts of about 3500 manuscript pages will be published as a standard-compliant, deeply annotated and highly integrated corpus under CC by-sa.[2]

While the project's goal is to enable an intensive exploration of Humboldt's lectures, our poster will abstract a little from the impressive range of topics of Humboldt's lectures and focus on technical solutions applied in the project to make the corpus likewise fruitful to Humboldt researchers, the general public and Dhists.

Starter

On March 13, 1828 Alexander von Humboldt gave his 14th talk of the Kosmos-Lectures at Berlin's Singakademie.[3] At the end of his talk he told his audience:

> … für den technischen Gebrauch hat man von diesen neueren Entdeckungen eine Anwendung gemacht, durch die Einrichtung eines *Ariometers*, od. Wollmessers. Ein Lichtstrahl welcher bei einem feinen Faden Wolle vorbei geht, erleidet eine Beugung u. bildet farbige Frangen oder

> Ringe, die um so breiter erscheinen als der Faden zarter od. dünner ist.[4]

The topic of this lecture is fairly clear: Optical interference and the diffraction of light. The main instrument mentioned is an “Ariometer” for measuring the thickness of woolen threads. This at least is the conclusion from what the anonymous writer of this text tells us.

Formal setting

The project’s tag set strictly follows the DTABf-M (cf. Haaf/Thomas, 2015)[5] which is a specific DTA “Base Format” (DTABf) for *manuscripts*. This strict subset of the TEI P5 tag set provides “a balance between expressiveness and precision as well as an interoperable annotation scheme for a large variety of text types in historical corpora […].” (Cf. Haaf et al., 2014)

The tag set and corresponding RNG-schema are the basis for a homogeneous TEI-encoding of all documents by different encoders, i. e. typists from the (external) vendor—providing transcription and TEI-conformant annotation of the easier legible manuscripts—as well as our project team.

The linguistic analysis of texts at the DTA (incl. lemmatization and fully automatized normalization of historical forms) and their search engine enable an ideal exploitation of the attendee’s lecture notes in the context of the DTA corpora, where hundreds of other texts from Humboldt himself as well as his contemporaries, predecessors and successors can be accessed and linked to the Kosmos-Lectures.

Obviously, as we are dealing with different notes by several attendees of Humboldt’s lectures—i. e. still referring to the same oral presentation—a collation of the witnesses is one essential way of access. The similarities and differences between the witnesses call for collation tools. CollateX or juXta, for instance, are well known and established DH tools for completely automatized detections and visualizations of comparisons of two or more texts. They are of great value for our project (cf. Thomas, 2014/15) and we will illustrate this by making accessible collation sets during our poster demo.

Due to our TEI-encoding more options for evaluations are given:

1. Humboldt’s course is segmented chronologically into single lectures.[6] This is thoroughly done via <div type="session" n="[count]">. This approach enhances the possibilities of accessing parallel reading of the transcripts.

2. A systematic access to the sources is enabled due to referencing persons within the TEI documents (<persName ref="[ID]">) as a basis for an encompassing index of all postscripts linked to authority records.

3. A third way of accessing the Kosmos-Lectures will be a complete bibliography of all mentioned sources tagged with <bibl>.

Contentwise Resumption

Starting point was the “Ariometer” and the diffraction of light in the one manuscript quoted above. A use case would be to see what the other listeners noted. For example, one other attendee wrote:

> … durch die Beugung der Lichtstralen, indem sie durch einen engen Raum gehn, entstehn farbige Franzen, die man am besten erhalten kann, wenn man einen Lichtstral durch das Fenster bei einem dünnen Faden vorbei in ein dunkles Zimmer auf eine helle Wand leitet…[7]

Similarities between the two texts are obvious: light beams, diffraction, a woolen thread and coloured stripes of light. But while the first extract gives the impression of Humboldt being interested in measuring woolen threads, these clearly function as instruments for illustrating a physical phenomenon in the later passage.

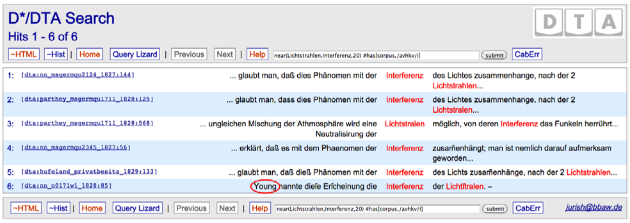
The natural attempt to search for “Ariometer” in our corpus gives no results.[8] We can derive from the context given in the quote above that Humboldt is talking about physical phenomena, but it would be helpful to have a protagonist. So let's refine our search:[9]

```
near(Lichtstrahlen,Interferenz,20) #has[corpus,/avhkv/i]
```

Looking at the results (see fig. 1) leads us to Thomas Young as well as Jung[10], which obviously only is a German version of the first. Next attempt:

```
near(*meter,/[JY]o?ung$/,20) #has[corpus,/avhkv/i]
```

And this leads to another *meter*, the “Circometer”. Maybe we couldn’t find the “Ariometer” since it doesn’t exist? To shorten the track here: A “Circometer” doesn’t exist either. What we are in fact looking for is the so-called *eriometer*.



D*/DTA Search
Hits 1 - 6 of 6

near(Lichtstrahlen,Interferenz,20) #has[corpus,/avhkv/i]

1:	[dta:nn_msgermqu2124_1827:144]	... glaubt man, daß dies Phänomen mit der	Interferenz	des Lichtes zusammenhange, nach der 2 Lichtstrahlen...
2:	[dta:parthey_msgermqu1711_1828:125]	... glaubt man, dass dies Phänomen mit der	Interferenz	des Lichtes zusammenhange, nach der 2 Lichtstralen...
3:	[dta:parthey_msgermqu1711_1828:568]	... ungleichen Mischung der Athmosphäre wird eine Neutralisirung der	Lichtstralen	möglich, von deren Interferenz das Funkeln herrührt...
4:	[dta:nn_msgermqu2345_1827:56]	... erklärt, daß es mit dem Phaenomen der	Interferenz	zusam̃enhängt; man ist nemlich darauf aufmerksam geworden...
5:	[dta:hufeland_privatbesitz_1829:133]	... glaubt man, daß dieß Phänomen mit der	Interferenz	des Lichts zusam̃enhänge, nach der 2 Lichtstrahlen...
6:	[dta:nn_n0171w1_1828:85]	Young nannte dieſe Erſcheinung die	Interferenz	der Lichtſtralen. –

near(Lichtstrahlen,Interferenz,20) #has[corpus,/avhkv/i]

jurish@bbaw.de

Fig. 1: Result of the DDC-SearchWhat we are facing here is an inherent difficulty of our text type: attendee’s lecture notes. It’s always possible that the listeners misheard something. This is why 1) the deeply granulated TEI encoding of the project, 2) the technical solutions provided by the DTA and 3) text-comparison visualizations (being developed for our project’s) are absolutely necessary.

Summary

The following DDC-search leads to the third volume of the Kosmos (1850), where Humboldt is talking about the above mentioned physical phenomena: near("$p=ADJA

/Fran[szg]en/",/(Wolle|Interferenz)/,20) #has[corpus,/(avhkv|kosmos)/i]

About 30 years after his lectures the eriometer and the wool are no issue any more. The reason for this is not inherent, but a little research on the basis of other digitized sources reveals it: Whereas the real eriometer—the apparatus for measuring the fineness of woolen threads—, which Young used to demonstrate the diffraction of light in his famous *Lectures on Natural Philosophy* at the Royal Institution, became forgotten,[11] the *idea* of combining the craft and the theory (or the engineer and the scientist) succeeded:

> … das ähnliche Instrument dient dazu die Güte der Wolle zu messen, und die Natur der Weltkörper zu bestimmen.[12]

Apart from the implementation of means for collation it is obvious that linking to more (external) data-sets, as shown here, will add even further value to the project.[13]

Bibliography

N. N. (1828). *Physikalische Geographie. Vorgetragen von Alexander von Humboldt*. In Deutsches Textarchiv, Berlin. http://www.deutschestextarchiv.de/nn_msgermqu2124_1827/(accessed 12 February 2016).

CollateX, http://collatex.net/; CollateX Console, http://collatex.net/demo/.

Erdmann, D. and Thomas, C. (2014). … zu den wunderlichsten Schlangen der Gelehrsamkeit zusammengegliedert. Neue Materialien zu den ‚Kosmos-Vorträgen' Alexander von Humboldts, nebst Vorüberlegungen zu deren digitaler Edition. *HiN – Humboldt im Netz. Internationale Zeitschrift für Humboldt-Studien*, **15**(28): 34–45, http://hin-online.de/hin28/erdmann-thomas.htm/ (accessed 12 February 2016).

George, S. and Guarino, M. (1973). Young's Eriometer: History and Modern Teaching Use. *Physics Education*, **8**(6): 392–96.

Haaf, S., Geyken, A. and Wiegand, F. (2014/15). The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources. *Journal of the Text Encoding Initiative (jTEI)*, **8**, http://jtei.revues.org/1114/ (accessed 12 February 2016).

Haaf, S. and Thomas, C. (2015). *DTABf-M: A TEI-conformant Base Format for Manuscripts.* Presentation at the TEI Converence and Members Meeting 2015, Lyon. http://tei2015.huma-num.fr/en/papers/#108 (accessed 12 February 2016).

Humboldt, A. V. (1845–62). *Kosmos. Entwurf einer physischen Weltbeschreibung*. 5 Bände. Stuttgart (u.a.): Cotta.

Juxta Collation Software, For Scholars, http://www.juxtasoftware.org/.

Parthey, G. (1828/28). *Alexander von Humboldt[:] Vorlesungen über physikalische Geographie*. Novmbr. 1827 bis April,[!] 1828. Nachgeschrieben von G. Partheÿ. Berlin, In Deutsches Textarchiv, http://www.deutschestextarchiv.de/parthey_msgermqu1711_1828/ (accessed 12 February 2016).

textloop Martina Gödel, http://textloop.de/.

Thomas, C. (2015). *Hidden Kosmos – Humboldts ‚Kosmos-Vorträge' als Probe der Digital Humanities*. Vortrag auf der DHd-Jahrestagung 2015 „Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation", 23.–27.2.2015, Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities an der Universität Graz. Book of Abstracts, http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege: [193]ff.

Young, T. (1807). *A Course of Lectures on Natural Philosophy and the Mechanical Arts*. Johnson.

Notes

[1] www.culture.hu-berlin.de/hidden-kosmos.

[2] https://creativecommons.org/licenses/by-sa/3.0/de/.

[3] Singakademie is todays Maxim-Gorki-Theater. After its opening on April 8th, 1827 it has been the biggest event hall in Berlin.

[4] http://www.deutschestextarchiv.de/nn_msgermqu2124_1827/14.

[5] More information at http://www.deutschestextarchiv.de/doku/basisformat_manuskripte.

[6] A contentual segmentation is in preparation.

[7] http://www.deutschestextarchiv.de/parthey_msgermqu1711_1828/126.

[8] If you give google search a try, this leads you to **Aräometer**, which is a totally different instrument. The reason for this is easy to tell: It's an OCR-mistake due to the "latin superscript small letter e".

[9] This and the following DDC searches are documented in detail here: http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe. In brief: We are looking for sentences including "Lichstrahlen" and "Interferenz" where there are no more than 20 tokens in between. And the search is restricted to our Kosmos-Lecture corpus (avhkv). If you want to give it a try, go to: http://www.deutschestextarchiv.de/search/ddc.

[10] Jung is not to be seen on the screenshot only because a little more context would be necessary. Just add \#cntxt2 to your search and choose the HTML-view.

[11] For "it seems hitherto to have been found much too delicate to be employed by the hard hands of peasants, with any advantage." (Cf. George/Guarino 1973, p. 392).

[12] http://www.deutschestextarchiv.de/nn_msgermqu2124_1827/147.

[13] The »ariometer« obviously is just one trace to follow. For an extended list of (the approx. 120 different) instruments mentioned please grab https://github.com/haoess/hidden-kosmos/blob/master/lists/Liste_Instrumente.txt.

Data-First Digital Humanities: How Adopting a Data-First Strategy Fosters Research, Collaboration, Pedagogy, and Scholarly Communication in the Digital Humanities

Todd Hughes
todd.hughes@vanderbilt.edu
Vanderbilt University, United States of America

Clifford Anderson
clifford.anderson@vanderbilt.edu
Vanderbilt University, United States of America

We propose a strategy for conducting digital humanities teaching and research that prioritizes publishing data above all other project activities. Drawing on our experience working with faculty, librarians, and graduate students on a critical edition in TEI of Charles Baudelaire's *Les Fleurs du Mal*, we demonstrate how adopting a data-first strategy fosters research, collaboration, pedagogy and scholarly communications in the digital humanities.

The *Corpus Baudelaire Project* began at Vanderbilt University in 2013, when a hybrid group of approximately ten scholars, who had recently learned how to encode literary texts in the TEI, aspired to do something practical with their new skills. The group developed a connection to Vanderbilt University Library's W. T. Bandy Center for Baudelaire and Modern French Studies; who exhaustively collects Baudelaire's works, including *Les Fleurs du Mal*. The work itself was published in four editions: 1857, 1861 (containing 35 additional poems, the *Tableaux parisiens*, and lacking six poems censored by the Second Empire), 1866 (including *Les Epauves* or *The Scraps*, and the six poems missing from the 1861 edition), and the posthumous 1886 edition. Participants in the *Corpus Baudelaire Project* are encoding all the editions using the critical edition apparatus in the TEI.

We describe our data-first approach to *Corpus Baudelaire Project*, which minimizes otherwise common tasks such as developing databases or coding interfaces, and argue for its advantage over alternative approaches in fostering collaboration, pedagogy, and new forms of publishing. We also suggest that our data-first approach may also productively be generalized to any digital humanities projects developing significant quantities of data.

A data-first approach differs from other forms of digital humanities scholarship by minimizing startup costs and reducing complexity. Whereas digital humanities projects aim above all to produce some form of online digital edition or interactive website, a data-first approach invests primarily in producing and sharing data with others. "It's the data, stupid!" is our informal slogan.

A data-first approach to DH involves at least three steps: licensing, curating, and publishing datasets online. The second two steps are likely to be iterative and emergent.

- **Licensing.** A data-first approach begins with the presupposition of making data openly available and reusable by other scholars. This not only implies attaching an open source license to the data, but also making certain that participants in the project understand the license and agree with its terms.
- **Curating**. A data-first approach implies that discussions about data curation start at the beginning of the project, not its end. How shall information be encoded? How to decide between alternative options? Are there emerging best practices and converging forms of representation? Documenting data and making available any accompanying schemas is also critical when taking a data-first approach.
- **Publishing**. A data-first approach requires that data be published for comment, criticism and reuse from the onset of the project. What are the best platforms for publishing digital humanities data? How can digital humanists provide access and get credit for their data?

By prioritizing these three activities above other forms of digital humanities, we simultaneously lower the barriers for participants to join our project while offering them the opportunity to publish and begin receive credit for their work almost immediately. Crucially, credit is allocated with respect to contributions, not by seniority or other hierarchical designations; the data bear witness directly to their creators.

Bibliography

Vishwas, Ch. and Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, 12.15 (2011): 1.

Rebuilding Digital Harlem for Sustainability and Change

Ian R. Johnson
ian.johnson@sydney.edu.au
University of Sydney, Australia

Artem Osmakov
osmakov@gmail.com
University of Sydney, Australia

Background

Digital Harlem (DigitalHarlem.org), winner of the *American Historical Association's* 2009 Roy Rosenzweig Prize for Innovation in Digital History, was a bespoke php/js/MySQL application with 34 tables and over 9,000 lines of code. While bespoke programming may under some circumstances offer the shortest path to a particular outcome, a fixed database structure and bespoke codebase pose problems for sustainability (the codebase will require ongoing maintenance and retention of support for code which is no longer current), for transferability (knowledge and development work which is not directly transferrable to other projects) and for evolutionary change (modification can involve significant rewriting and programming expertise). For Digital Harlem both evolving requirements and external changes have forced the excavation and relearning of code long after development funding ceased, an experience which will be common for any project which seeks longevity beyond short-term project grants.

Conversion

In 2014/2015 we converted the Digital Harlem database to Heurist (HeuristNetwork.org). This allowed us to unlock the previously inflexible data structures and rigid interface, and refine the data model, enabling the project to start a campaign of data entry for a new research grant and focus. Where analysis of the data was previously restricted to three rather limiting search forms for People, Places and Events, the new version (Figure 1) opens up a full range of built-in data management functions and user-defined searches/filters, including multi-level faceted search. Search results can now be saved and visualised with maps, timelines, network diagrams and user-defined reports, as well as file and printed output.

The original public website was subsequently reimplemented, with minor external changes, as a reskinned view of the database running natively within Heurist (Figure 2). We moved significant elements of the interface - search forms, base maps, popup content, buttons – out of custom code into data. These data can be easily edited by the research team allowing them to extend the interface without technical assistance. Fixed form-based searches were replaced with saved faceted searches which can be added to or modified without programming.

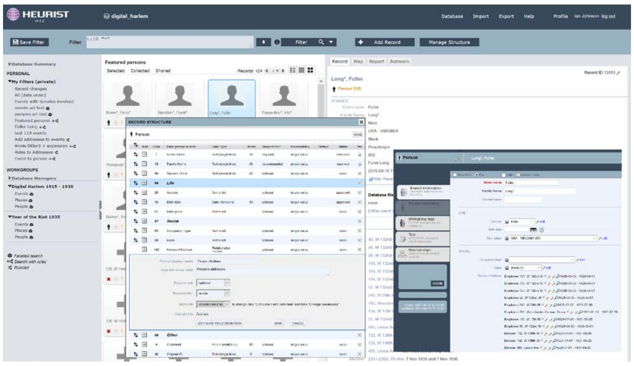


Figure 1. Digital Harlem - the standard Heurist interface (used by the research team)

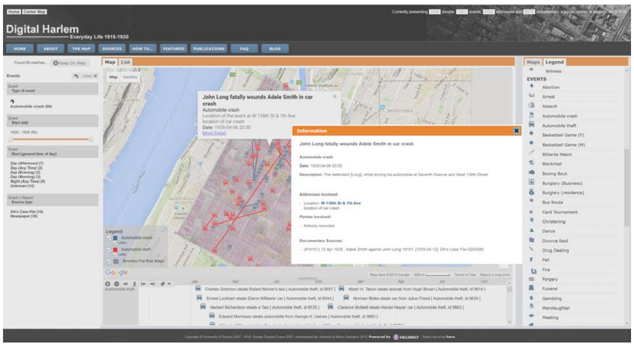


Figure 2. Digital Harlem - the reimplemented public interface with faceted search

Adaptability

The public interface is easily adaptable to other projects requiring a customised public search, mapping and timeline interface for richly linked entities. It is not tied to specific types of entity or relationship, as almost all customisation other than visual appearance occurs within the database content. The interface is built from reusable widgets in a responsive framework, using less than 1,000 lines of html, css, php and js code. New widgets can be added for additional types of interaction, although many projects will find the existing widgets adequate.

Sustainability

Heurist databases retain the inherent medium-term sustainability of an Open Source MySQL database at the backend, but reinforce this sustainability through the adoption of an identical structure across many diverse projects. The use of a single, well-documented database structure across all projects promotes the transfer of expertise and leverages the effort of code development - when someone requests Zotero synchronisation or the generation of GEPHI network files, the code can be written just once and every database inherits the capability. The same goes for any maintenance required to keep pace with the changing web environment and for bug fixing. Standard documented SQL queries, which can be run from any programming language, will work across all databases. A complete, fully documented XML data archive can be generated in a couple of clicks.

Conclusion

In this poster we will outline the sustainability and development benefits of the new implementation of Digital Harlem. By adopting an adaptable codebase (Heurist) which can run many heterogeneous projects we have leveraged development effort and benefits across Digital Harlem and several dozen other projects. For new projects, common data structures can be imported with a few mouse clicks from a clearinghouse of projects, adapted to specific needs and republished for use by others. Existing public interfaces can, with slightly more effort, be repurposed for new projects. A stable well-documented underlying data format also allows independent code development in a variety of languages.

If there is a final takeaway, it is that - while there will always be a place for bespoke, one-off code as a vehicle for experimentation - for the majority of projects, re-use of a shared codebase and body of expertise (where practical) will be more cost-effective, require less technical development and provide a better chance of longevity. Heurist offers one such generic solution which has allowed Digital Harlem to escape its self-imposed straightjacket of bespoke data structure and code; new data structures, research tools or public interfaces can simply be added without reengineering the system. Digial Harlem builds on the work of many preceding projects, and future projects can build on this experience without the cost of reinvention.

Digital Resources and Research Data in the Digital Humanities: The Digital Knowledge Store at the Berlin-Brandenburg Academy of Sciences and Humanities

Marco Jürgens
juergens@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Sascha Grabsch
grabsch@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

The Digital Knowledge Store was developed from March 2012 to April 2015 at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and funded by the Deutsche Forschungsgemeinschaft (DFG). In this first project phase a search infrastructure enables centralised access to all digital resources of the BBAW was created. The DFG granted funds for a second period in which the Digital Knowledge Store will be expanded with a focus on deployment options and integration for partner research institutions.

The BBAW hosts over 170 research projects with over 1.2 million digital resources. For the first time these resources and their metadata were completely integrated into a central full text index and made accessible through an innovative user interface. The resources hosted at the BBAW vary widely in terms of content, formats and languages. The main part of the resources are provided as digital editions and translations in formats like XML, HTML and PDF but also as electronic catalogues, documentations, databases, digital full text collections and dictionaries. The Digital Knowledge Store can be queried in different languages via the morphologically analyzed index. We utilize a number of language technologies (Bing, DONATUS) to enable this multilingual search. The search also covers automatically and manually created metadata which enhance the resources, connect them semantically and provide additional information to the user. The metadata of all resources are provided as well via a machine readable web service (OAI-PMH) and in that way become part of the Linked Open Data Cloud.

The biggest challenge in building an interdisciplinary research data infrastructure like the Digital Knowledge Store was the heterogeneity of the digital resources created at the BBAW in the last 20 years. Hosted on different servers in different databases they vary widely in regard to content and access possibilities. It was the main task to access these data generically and bundle them in the central Apache Lucene index and in a metadata scheme adapted to the needs of the academy (based on OAI-ORE). Specific import modules were implemented for the various projects and resource collections which integrate the varying data structures of the research projects. Semantically connected suggestions are provided by integrating Semantic Web Technologies (e.g. DBpedia) and Text Mining components which extend the query term and invite the user to discover and explore the academy's projects.

The second project phase of the Digital Knowledge Store running until 2017 will expand its possibilities especially in terms of sustainability and availability. There is a heavy demand by academic institutions for sustainable longterm research infrastructures which can meet the specific requirements of research data in the humanities, e.g. the integration of heterogeneous resources and content handling. One important goal in the next stage is to broaden the target user group. The software components of the Digital Knowledge Store will be provided as an installation package. This will enable Partner institutions to run their own Knowledge Store adapted to their own digital resources. In order to coordinate further development and collaboration with future users an open workshop will be held in April 2016 in Berlin.

Another topic in the next project period will be the development of guidelines for the minimum structural and technical requirements that resources and metadata have to meet to be integrated easily into the index. The guidelines will include objectives for the (technical) quality of the resources and their metadata. These best-practice-recommendations can become a general recommendation in the digital humanities for building and maintaining resource collections and a reference on how to deal with the quality of resources and metadata beyond their specific use case. Our partner institutions will successively optimize and adjust them to their specific needs. Additionally workflows for the manual and automatic supply of metadata will be created and specified. Further development goals are the automatic evaluation and integration of user feedback into the query process as well as visualization components.

Bibliography

Ballsun-Stanton, B. (2012). *Asking About Data - Exploring Different Realities of Data via the Social Data Flow Network Methodology.* The University of New South Wales. http://www.mendeley.com/download/public/2110651/4867189482/58d704a31071163071cb0a391f17ad202fe958ff/dl.pdf (accessed 17 October 2015).

Ballsun-Stanton, B. (2009). Philosophy of Data (PoD) and Its Importance to the Discipline of Information Systems, *AMCIS 2009 Proceedings.* Paper 435. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1443&context=amcis2009 (accessed 17 October 2015).

Burdick, A. (2012). *Digital Humanities*. Cambridge: Cambridge University Press.

De Roure, D. (2011). Machines, Methods and Music: On the Evolution of e-Research. In *High Performance Computing & Simulation,* Oxford. http://users.ox.ac.uk/ oerc0033/preprints/hpcs11.pdf (accessed 17 October 2015), pp. 8–13.

Dörk, M. et al. (2011). The information flaneur: a fresh look at information seeking. In *CHI 2011. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 1215-24.

Fensel, D. (2004). *Ontologies. A Silver Bullett for Knowledge Management and Electronic Commerce.* Springer-Verlag Berlin Heidelberg.

Franklin, M., Halevy, A. and Maier, D. (2005). From databases to dataspaces: a new abstraction for information management. *ACM Sigmod*, **34**(4): 27–33.

Haffner, A. (2012). *Internationalisierung der GND durch das Semantic Web.* http://www.kim-forum.org/Subsites/kim/SharedDocs/Downloads/DE/Berichte/internationalisierungDerGndDurchDasSemanticWeb.pdf?__blob=publicationFile (accessed 17 October 2015).

Jannidis, F. (2010). Methoden der computergestützten Textanalyse. In Nünning, V. and Nünning A. (eds), *Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze - Grundlagen – Modellanalysen.* Stuttgart, pp. 109–32.

Unsworth, J. (2011). Computational Work with Very Large Text Collections. *Journal of the Text Encoding Initiative*, **1**: 1-9.

Voß, J. (2013). *Describing Data Patterns - A general deconstruction of metadata standards.* Pd.D thesis, Humboldt-University Berlin, http://edoc.hu-berlin.de/dissertationen/voss-jakob-2013-05-31/PDF/voss.pdf (accessed 4 March 2016).

Voß, J. (2013). Was sind eigentlich Daten? In *LIBREAS. Library Ideas No 23.* http://libreas.eu/ausgabe23/02voss/ (accessed 17 October 2014).

W3C Working Group (2014). RDF 1.1 Primer. http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/ (accessed 17 October 2015).

Ward, D., Hahn, J. and Feist, K. (2012). Autocomplete as Research Tool: *A Study on Providing Search Suggestions. Information Technology and Libraries*, **31**(4): 6-19.

Whitelaw, C., Hutchinson, B., Chung, G. and Ellis, G. (2009). Using the Web for Language Independent Spellchecking and Autocorrection. In *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, **2**: 890-99.

Whitelaw, M. (2012). Towards Generous Interfaces for Archival Collections. *Paper for International Council on Archives Congress.* http://mtchl.net/wordpress/wp-content/uploads/2013/10/Whitelaw_ICA_GenerousInterfaces.pdf (accessed 17 October 2015).

Visual Forms of Information Presentation and Their Place in Formal Digital Scientific Communication

Anna Kamińska
a.kaminska8@uw.edu.pl
University of Warsaw, Poland

The purpose of this paper is answering the question concerning the place of the visual transfer in the formal system of digital scientific communication. The question is important for scientists who publish or would like to publish their works in a visual way, because the question is connected with issue of the evaluation of the scientific achievements. The bibliometric impacts, like Impact Factor or Hirsch Index, respect only the publications that take part in the formal scientific communication. Attendance or absence of visual forms of information presentation in the formal scientific communication might influence the scientist's decision about presenting scientific works in a visual way and thereby popularise the idea of digital humanities.

The formal scientific communication is difficult to unambiguously define the cause of the fuzzy limit between formal and informal scientific communication. In general, before the publications become part of the formal scientific communication, they should be edited, reviewed, published and indexed in bibliography, bibliographic databases, scientific search engines (Nahotko, 2010; Pikas, 2006; Sapa, 2009). In this research the visual forms of information presentation were analysed using those four features of the formal scientific communication. Scientific papers, communication channels and bibliographies had to be digital and be accessible by Internet.

The visual forms of information presentation include forms using static or dynamic pictures as information media (maps, diagrams, infographics, 3D animations and movies) and textual digital publications whose design and navigation were strongly visual like *The Roaring 'Twenties: an interactive exploration of the historical soundscape of New York City* (Thompson and Mahoy, 2013). The visual forms of information presentation had to be about humanities topics.

Forty-one English-speaking digital humanities projects were collected by the Polish scientific literature and DH Awards project ("Digital Humanities Awards", n.d.). It is difficult to state what proportion of all digital humanities projects constitute the collected projects because of insufficient registration. Therefore, the generalisation of the research results is impossible. However, the analysis of the collected projects enables to notice problems in presenting information in a visual way with digital formal scientific communication and offer some solutions.

It was found that maps, graphs, movies and animations were most frequently published on websites connected with some specific projects of digital humanities, institutions or initiatives devoted to digital humanities, carrying out many projects. Very occasionally, they were found on scientific blogs, academic social networks such as academia.edu and general portals like Flick, YouTube or Vimeo. The analysis of the visual forms of information presentation proved that the collected material had not been previously edited almost in all cases. It seemed that the reverse tendency would be shown in case of websites on which the projects were published. Still, only few websites contained information concerning their editors. No information on the review of the visual forms of information presentation were found. After considering the place where they were published it turned out that the information including its review were found only in the multimedia "Vectors Journal".

Since the leading interdisciplinary bibliographic databases (like Web of Science and Scopus), subject bibliographic databases (like MLA International Bibliography) and scientific search engines (like Google Scholar) register only texts – the journal articles, the books or the conference proceedings, searching for visual forms of information presentation in these databases seemed to be pointless.In conclusion, the lack of review, unsatisfactory registration in bibliographic databases and scientific search engines were observed as the main problems for visual forms of information presentation to become part of digital formal scientific communication. In this place it is worth asking what should be registered – the movie, the animation, the diagram etc. or the place, where they were published (like website or multimedia publication). Not all visual forms of information presentation could be recognized as individual publications.

The author of this paper proposes some solutions how visual forms of information presentation might become part of digital formal scientific communication. Visual forms of information presentation or places where visual forms are published could provide ISBN, ISSN or DOI number and place the bibliographic data. Bibliographic databases and scientific search engines should change the current rules of registration.

Bibliography

Bomba, R. (2013). Narzędzia cyfrowe jako wyznacznik nowego paradygmatu badań humanistycznych. In Radomski, A. and Bomba, R. (eds), *Zwrot Cyfrowy W Humanistyce*. Lublin: E-naukowiec, pp. 57–71. http://depot.ceon.pl/handle/123456789/3128 (accessed 5 March 2016).

City Witness Project. http://www.medievalswansea.ac.uk/en/ (accessed 8 May 2015).

Digital Humanities Awards. http://dhawards.org/ (accessed 4 May 2015).

DOI Foundation (2015). *DOI Handbook*. http://www.doi.org/doi_handbook/TOC.html.

Drucker, J. (2012). Humanistic theory and digital scholarship. In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minneapolis: University Of Minnesota Press, pp. 85–95.

Gmiterek, G. (2014). Książka w erze nowych technologii, integracji i interaktywności mediów. In Sobczak, A., Cichocka, M. and Frąckowiak, P. (eds), *Historia 2.0 : Panta Rhei Materiały Sympozjum XIX Powszechnego Zjazdu Historyków Polskich 17 Września 2014 W Szczecinie*. Lublin: E-naukowiec, pp. 67–74. https://repozytorium.lectorium.pl/handle/item/898 (accessed 5 March 2016).

International ISBN Agency (2012). *ISBN Users' Manual*. 6th ed. London: International ISBN Agency. https://www.isbn-international.org/sites/default/files/ISBN Manual 2012 -corr.pdf (accessed 5 March 2016).

International Standard Serial Number International Center (2005). *ISSN Manual: International Standard Serial Number*. http://www.issn.org/wp-content/uploads/2013/09/ISSN-Manual_ENG2015_23-01-2015.pdf (accessed 5 March 2016).

Jessop, M. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing*, **23**(3): 281–93.

Kindred Britain. http://kindred.stanford.edu/# (accessed 10 May 2015).

Mapping the Republic of Letters. http://republicofletters.stanford.edu/index.html (accessed 8 May 2015).

Nahotko, M. (2010). *Komunikacja Naukowa W środowisku Cyfrowym : Globalna Biblioteka Cyfrowa W Informatycznej Infrastrukturze Nauki*. Warszawa: Wydawnictwo SBP.

ORBIS: The Stanford Geospatial Network Model of the Roman World. http://orbis.stanford.edu/ (accessed 10 May 2015).

Pikas, C. K. (2006). *The Impact of Information and Communication Technologies on Informal Scholarly Scientific Communication: A Literature Review*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.9216&rep=rep1&type=pdf (accessed 5 March 2016).

Radomski, A. (2014). *Humanistyka W świecie Informacjonalizmu*. http://e-naukowiec.eu/wp-content/uploads/2014/06/A.Radomski.pdf (accessed 5 March 2016).

Sapa, R. (2009). *Metodologia Badań Obszaru Pośredniczenia W Komunikacji Naukowej Z Perspektywy Nauki O Informacji*. Kraków: Wydawn. Uniwersytetu Jagiellońskego.

Słownik języka polskiego PWN. http://sjp.pwn.pl/sjp/wizualny;2579950.html (accessed 3 May 2015).

Terras, M. Infographic: Quantifying Digital Humanities. http://melissaterras.blogspot.co.uk/2012/01/infographic-quanitifying-digital.html (accessed 6 May 2015).

The Valley of the Shadow: Two Communities in the American Civil War. http://valley.lib.virginia.edu/ (accessed 5 March 2016).

Thompson, E. and Mahoy, S. (2013). The Roaring 'Twenties : an interactive exploration of the historical soundscape of New York City. *Vector: Journal of Culture and Technology in a Dynamic Vernacular*, **4**(1). http://vectorsdev.usc.edu/NYCsound/777b.html (accessed 5 March 2016).

Vectors : Journal of Culture and Technology in a Dynamic Vernacular. http://vectors.usc.edu/journal/index.php?page=Introduction (accessed 5 May 2015).

Virtual Pauls Cross Website: a Digital re-creation of John Donne's Gunpowder Day sermon. http://vpcp.chass.ncsu.edu/ (accessed 19 May 2015).

Wieczorek-Tomaszewska, M. (2013). Cyfrowa humanistyka jako metaforyczna współczesna Republika Listów. *23. Ogólnopolskie Sympozjum Naukowe „Człowiek - Media - Edukacja" 27-28 Września 2013*. Kraków. http://ktime.up.krakow.pl/symp2013/referaty_2013_10/wieczorek.pdf (accessed 5 March 2016).

Wilkowski, M. (2013). *Wprowadzenie Do Historii Cyfrowej*. Gdańsk: Instytut Kultury Miejskiej. https://depot.ceon.pl/handle/123456789/2001 (accessed 5 March 2016).

The Royal Society Corpus: Towards a high-quality corpus for studying diachronic variation in scientific writing

Hannah Kermes
h.kermes@mx.uni-saarland.de
Universität des Saarlandes, Germany

Stefania Degaetano-Ortlieb
s.degaetano@mx.uni-saarland.de
Universität des Saarlandes, Germany

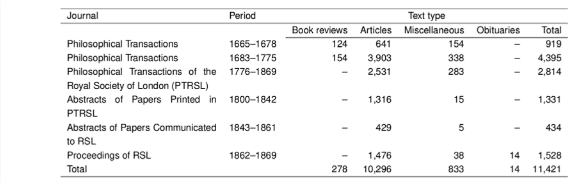
Ashraf Khamis
a.khamis@uni-saarland.de
Universität des Saarlandes, Germany

Jörg Knappen
j.knappen@mx.uni-saarland.de
Universität des Saarlandes, Germany

Elke Teich
e.teich@mx.uni-saarland.de
Universität des Saarlandes, Germany

Big data are a potential source for quantitative research in the humanities, but typically they do not contain all relevant contextual meta-data (time, register/genre, social group, author, etc.) to be readily usable for social, historical or philological studies (cf. Schöch, 2013). Small corpora, in contrast, are typically carefully hand-crafted and provide rich meta-data as well as structural and linguistic data, but the application of data-driven analysis techniques is impeded by their small size.

We introduce a diachronic corpus of English scientific writing - the Royal Society Corpus (RSC) - adopting a middle ground between big and 'poor' and small and 'rich' data. The corpus has been built from an electronic version of the Transactions and Proceedings of the Royal Society of London and comprises c. 35 million tokens from the period 1665- 1869 (see Table 1). The motivation for building a corpus from this material is to investigate the diachronic development of written scientific English.



Journal	Period	Text type				
		Book reviews	Articles	Miscellaneous	Obituaries	Total
Philosophical Transactions	1665–1678	124	641	154	–	919
Philosophical Transactions	1683–1775	154	3,903	338	–	4,395
Philosophical Transactions of the Royal Society of London (PTRSL)	1776–1869	–	2,531	283	–	2,814
Abstracts of Papers Printed in PTRSL	1800–1842	–	1,316	15	–	1,331
Abstracts of Papers Communicated to RSL	1843–1861	–	429	5	–	434
Proceedings of RSL	1862–1869	–	1,476	38	14	1,528
Total		278	10,296	833	14	11,421

Table 1: Material used for the RSC

In terms of corpus building (see Figure 1 for a schematic overview), the sources for the RSC were obtained from JSTOR and include some but not all relevant meta-data (year of publication and authors, but not disciplines), structural data is partial and erroneous (e.g. scrambled pages, text duplicates), and the base text contains OCR errors. To move towards a cleaner and richer version of the corpus, an approach is needed that allows obtaining good-quality base-text data and relevant meta-data as well as structural and linguistic data with affordable effort. For this purpose, we use a combination of pattern-based techniques (e.g. by adapting the patterns for OCR corrections made available by Underwood and Auvil)[1] and data-mining methods (e.g. topic modelling (Blei et al.,2003)to approximate disciplines; cf. McFarland et al.(2013)for an overview of types of topic models applied to capture differentiation in scientific language). Additionally, to enrich the RSC with basic linguistic annotations, we build on existing tools adapting them to the diachronic material. For normalization we use VARD (Baron and Rayson,2008)with a model we trained on a manually normalized subset of the RSC, and for tokenization, lemmatization, segmentation and part-of-speech annotation we useTreeTagger (Schmid,1994)on the normalized texts. Inspired by the idea of Agile Software Development (Cockburn, 2001), we intertwine the actual corpus building with corpus annotation and analysis, continuously building new versions of the corpus whenever we see a recurrent problem in data quality. We work with a dedicated pipeline and keep the corpus-building process as modular and automatic as possible, applying manual work before the first automatic step. In the last step, the corpus is encoded in CQP format (cf. IMS Open Corpus Workbench (CWB) (Evert and Hardie, 2011)) and can be accessed via a CQPweb interface (Hardie, 2012)[2].

In terms of analysis, our main assumption is that due to specialization, scientific texts exhibit greater encoding density over time (Halliday and Martin,1993),resulting in a specific discourse type characterized by high information density (Crocker et al., 2015) that is functional for expert communication (but rather inaccessible to lay persons). Linguistically, this may be reflected in lexical compression (e.g. compounding, derivation) and syntactic reduction (e.g. relativizer omission, contractions). For instance, there is evidence from the Thesaurus of the OED (Oxford English Dictionary)[3] that affixation rises considerably as a means of word formation in scientific texts in the mid-17th century. For the identification that affixation rises considerably as a means of word formation in scientific texts in the mid-17th century. For the identification of further linguistic features possibly involved in denser encoding, we draw, on the one hand, on existing literature (e.g. Harris, 1991) and, on the other hand, on exploratory data-mining techniques (e.g. pattern mining as in Vreeken, 2010).

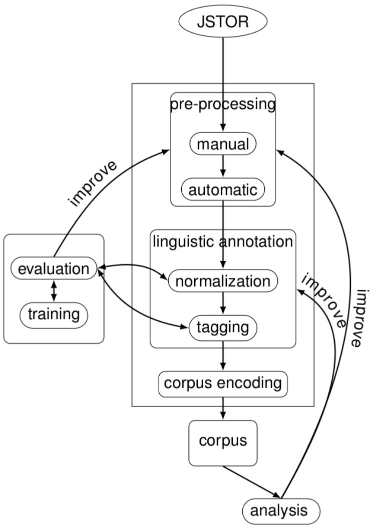


Figure 1: Corpus building steps

In the poster, we show the corpus-building process and selected analyses of diachronic development in the RSC with dedicated visualizations (Fankhauser et al., 2014).

Bibliography

Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings ofthePostgraduateConferenceinCorpusLinguistics*, Birmingham, UK.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**: 993–1022.

Cockburn, A. (2001). *Agile Software Development*. Addison-Wesley Professional, Boston, USA.

Crocker, M. W., Demberg, V. and Teich, E. (2015). Information density and linguistic encoding (IDeaL). *KI - Künstliche Intelligenz*, pp. 1–5.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics Conference*, Birmingham, UK.

Fankhauser, P., Kermes, H. and Teich, E. (2014). Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity. In *Digital Humanities*, Lausanne, Switzerland.

Halliday, M. & Martin, J. (1993). *Writing science: literacy and discursive power*. Falmer Press, London.

Hardie, A. (2012). *International Journal of Corpus Linguistics*

17(3): 380-409.CQPweb – combining power, flexibility and usability in a corpus analysis tool.

Harris, Z. S. (1991). *A theory of language and information: a mathematical approach*. Oxford University Press, USA.

McFarland, D.A., Ramage, D., Chuang, J., Heer, J., Manning, C.D. and Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, **41**(6): 607–25.

Notes

[1] http://usesofscale.com/gritty-details/basic-ocr-correction/

[2] https://fedora.clarin-d.uni-saarland.de/cqpweb/

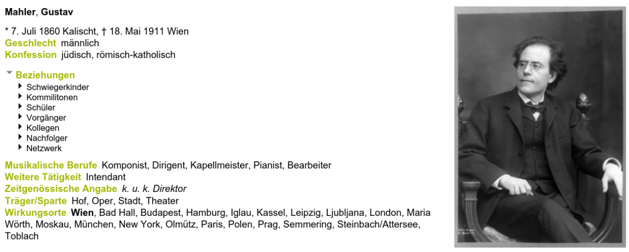
[3] http://www.oed.com/thesaurus/

On the Distant Reading of Musicians' Biographies

Richard Khulusi
richard.khulusi@web.de
Leipzig University, Germany

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

The Bavarian Musicians Encyclopedia Online (Bayerisches Musiker Lexikon Online, BMLO) is a web-based platform that provides access to biographical information about musicians associated to Bavaria's music history (BMLO, 2016). Most of the musicians contained in the corresponding database had an active lifetime period living in Bavaria or a considerable influence on Bavaria. Initiated in 2004, the musicians database contains biographical data about nearly 28,000 musicians now. This suggests the rather global scope of the BMLO – underpinned by many musicologists worldwide using the BMLO for their daily work. The screenshot in Figure 1 shows the BMLO entry for the composer Gustav Mahler.



Mahler, Gustav

* 7. Juli 1860 Kalischt, † 18. Mai 1911 Wien
Geschlecht männlich
Konfession jüdisch, römisch-katholisch

Beziehungen
- Schwiegerkinder
- Kommilitonen
- Schüler
- Vorgänger
- Kollegen
- Nachfolger
- Netzwerk

Musikalische Berufe Komponist, Dirigent, Kapellmeister, Pianist, Bearbeiter
Weitere Tätigkeit Intendant
Zeitgenössische Angabe *k. u. k. Direktor*
Träger/Sparte Hof, Oper, Stadt, Theater
Wirkungsorte **Wien**, Bad Hall, Budapest, Hamburg, Iglau, Kassel, Leipzig, Ljubljana, London, Maria Wörth, Moskau, München, New York, Olmütz, Paris, Polen, Prag, Semmering, Steinbach/Attersee, Toblach

Figure 1: Biographical information about Gustav Mahler in the BMLO

A recently published article facilitates the profiling of musicians based on the BMLO (Jänicke et al., 2015). The profile of a musician of interest can be visualized, and according to biographical information, similar musicians are determined in a semi-automated process (MusikerProfiling, 2016) . A profiling result for Gustav Mahler is shown in Figure 2. Although the profiling system has been proven useful for the collaborating musicologists, it does not support generic research questions like "In which cities Roman Catholic conductors worked during the 18th century?" or "What are the differences and similarities among the careers of pianists and violinists?" Therefore, the musicologists desired a system that facilitates the dynamic exploration of musicians' characteristics with the help of interactive visual interfaces. The design of the resultant visualization system is outlined below.

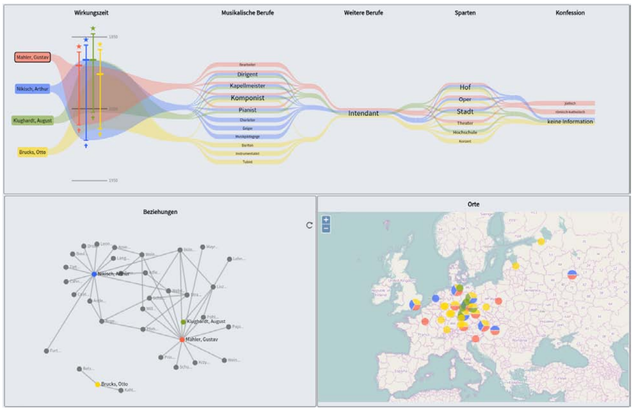


Figure 2: Interactive visual profiling of Gustav Mahler comparatively visualizes Mahler's profile to the profiles of the three most similar musicians in three views (Column Explorer, Relationship Graph, Map)

To support the dynamic exploration of musicians' biographies, we provide various views that visualize aggregate biographical information of musicians inherent in the database. For the divisions where musicians worked, we use a tag cloud (Fig. 3a). As musical (Fig. 3b) and further professions (Fig. 3c) are organized in a hierarchy, we apply a sunburst technique tailored for such structures (Stasko et al., 2000). A map plots all places of activity (Fig. 3d). Using GeoTemCo (Jänicke et al., 2013) for that purpose, occluding dots are clustered and metropolises of music history, e.g., Munich, Vienna and Berlin, are salient as large circles. To illustrate the denominations of musicians, we use again a tag cloud (Fig. 3e), and a pie chart to visualize musicians' sexes (Fig. 3f). Finally – based on the dates of birth, the first mentioned dates and the dates of death – we define an activity time for each musician. The aggregate of all activity times is shown in a timeline graph (Fig. 3g). With mouse interaction, each view can be used for filtering purposes. So, the investigation of rather generic research questions in musicology gets possible.

Figure 4 shows the filter steps required to explore in which cities Roman Catholic conductors worked during the 18th century. The first filter is applied in the denomination tag cloud (Fig. 4a) by clicking "römisch-katholisch"

(Roman Catholic). Then, we select "Kapellmeister" (conductor) in the musical profession sunburst plot (Fig. 4b). Finally, dragging a time range from 1700 to 1800 in the timeline (Fig. 4c) filters the dataset from around 28,000 to 72 musicians, and their places of activity get visible in the map.

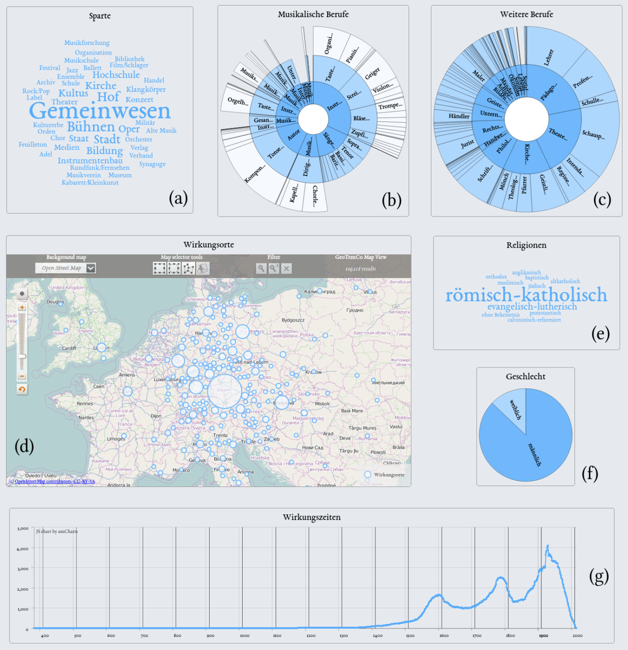


Figure 3: Seven visual interfaces to explore various musicians' characteristics

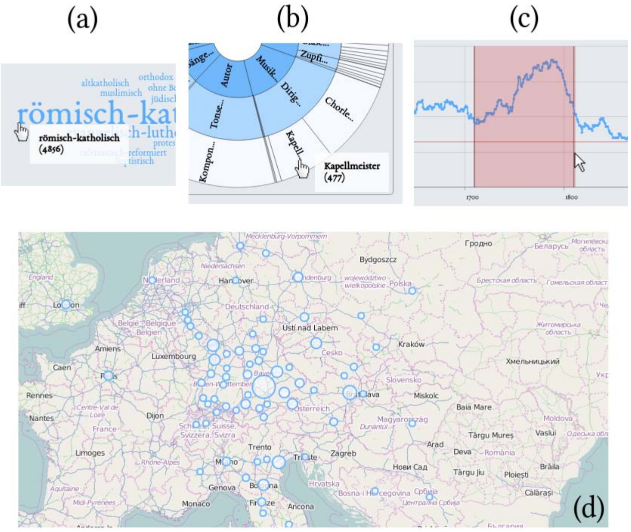


Figure 4: Exploring the question "In which cities Roman Catholic conductors worked during the 18th century?"

The system also supports the comparison of musicians groups. An example is given in Figure 5, which investigates the differences and similarities among the careers of pianists and violinists . We are using different colors to mark the attributes of violinists (yellow) and pianists (blue). The several views are slightly modified to support the comparative analysis. Looking at the result, some initial conclusions can be drawn:

- Although the database contains more violinists (Geiger) than pianists (Pianist), the profession of a pianist seems to be more multifarious, e.g., pianists had more musical and more further professions than violinists.
- The pianist profession is newer than the violinist profession.
- The ratio between male and female musicians is more balanced for pianists.
- Whereas Munich is the city where lots of pianists worked, violinists worked often in northern Bavaria, e.g., Bayreuth (772 x).

In our poster presentation, we will illustrate the above outlined design of the visualization system. In addition, we would demonstrate typical usage scenarios from the collaborating musicologists.

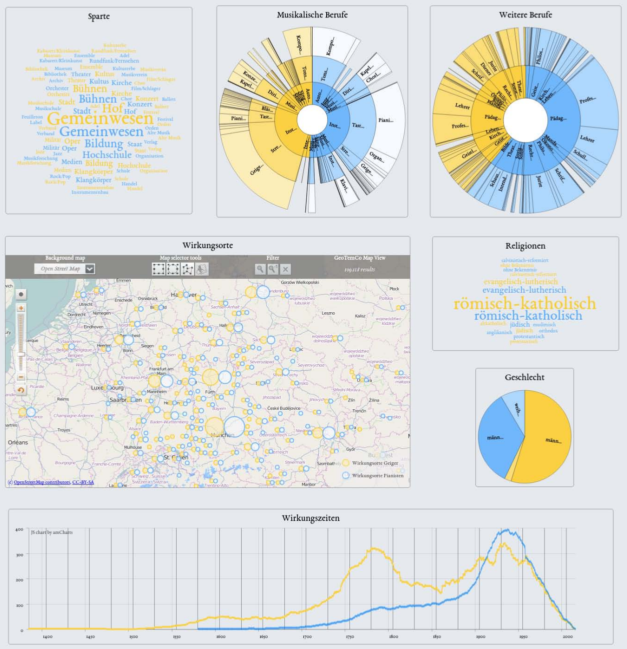


Figure 5: Exploring the question "What are the differences and similarities among the careers of pianists and violinists?"

Bibliography

BMLO (2016). Bayerisches Musiker-Lexikon Online, 2016. ed. Josef Focht. http://www.bmlo.lmu.de/ (accessed 2 March 2016).

Jänicke, S., Focht, J. and Scheuermann, G. (2016). Interactive Visual Profiling of Musicians. *Visualization and Computer Graphics, IEEE Transactions on,* **22**(1): 200–209, Jan 2016.

Jänicke, S., Heine, C. and Scheuermann, G. (2013). GeoTemCo: Comparative Visualization of Geospatial-Temporal Data with Clutter Removal Based on Dynamic Delaunay Triangulations. In *Computer Vision, Imaging and Computer Graphics. Theory and Application,* pp. 160–75.

MusikerProfiling (2016). Musiker Profiling der Universität Leipzig. http://profiling-musicians.vizcovery.org/ (accessed 2 March 2016).

Stasko, J. and Zhang, E. (2000). Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on,* pp. 57-65.

The Encoded Medieval Antiphoner: an Open Access Digital Source for Music and Liturgical Pedagogy and Scholarship

Anna Ewelina Kijas
anna.kijas@bc.edu
Boston College, United States of America

This poster will provide an overview of the development of an open-access digital musicology project, *The Encoded Medieval Antiphoner: an Open Access Digital Source for Music and Liturgical Scholarship* at Boston College. This is a collaborative project between the Digital Scholarship Group and library staff at the Boston College University Libraries, musicologist Dr. Michael Noone, research assistants, and several external partners, including CANTUS database staff. In the summer of 2015 we began encoding a 14th century Franciscan Antiphoner using the Music Encoding Initiative (MEI), as well as developing an open-access site to present the manuscript as an interactive object for research and scholarly use. The original Antiphoner is in manuscript form bound between leather-covered boards containing 119 parchment folios with text and notation for antiphons and responsories for the entire annual calendar of saints' days (sanctorale). The main goals of this project will be presented in this poster. These include:

1. Making the Antiphoner interactive by building an open access and responsive website with a dynamic presentation layer that will include searchable content (metadata, notation, XML/MEI) and multimedia;
2. Using and developing open source technology to encourage and support further development in the open source community, as well as, sharing of documentation and data;
3. Contributing our data to the scholarly community through a collaboration with CANTUS, a database for Latin Ecclesiastical Chant; and
4. Developing software, workflows, and documentation that can inform future projects using similar technology.

This poster will explain our process and workflows, which have involved transcribing and encoding over 1500 musical incipits, texts, and metadata, contributing the data to the CANTUS database, a highly-respected digital archive and index of chants, as well as, implementing open-source software for a presentation layer and search/retrieval of the content. The digitized object will be presented using Diva.js an open-source software that connects with the API of the CANTUS database to pull in metadata, transcribed music notation, and bibliographic data that we have contributed to this database. We will also present a variety of functionality options that will make this Antiphoner interactive. Basic functionality, which is enabled in Diva.js, includes the ability to:

- view the manuscript side-by-side with metadata from the CANTUS database;
- run search queries on both the text and notation in the manuscript;
- select viewing options: pages (book), individual page, thumbnail/gallery view;
- zoom in/out; full-page view
- access a permanent link/URL for each manuscript page

Beyond the basic functionality listed above, we may explore additional features, such as the ability to display manuscript pages in non-chronological order side by side for comparison and analysis, or allow users to annotate, add metadata, and manipulate content within the manuscript through a framework like the Shared Canvas Data Model.

Additionally, this poster will highlight the collaborative aspect of this project and several positive outcomes, such as an interest to investigate rendering neume notation using Verovio rather than Volpiano font. Verovio is a software that renders MEI directly in a modern browser as SVG (scalable vector graphics: XML-based vector image format), however at this time it can only render MEI files that use Common Western Music Notation (CMN) elements and attributes, not those associated with the Neume module. Our combined interest and needs will allow us to explore this option that can benefit not only our project, but could enable scholars to contribute MEI files with their data into CANTUS and have their notation rendered using Verovio. This could potentially expand the mission of CANTUS Database to also provide access to MEI/XML data.

Bibliography

John J. *Franciscan Antiphoner (sanctorale)*. Burns Library, Boston College, University Libraries. http://hdl.handle.net/2345/2231.

Moving Images and the Connection to other Media Types

André Kilchenmann
a.kilchenmann@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

In film and media studies, there has always been a desire to annotate and analyze movies as easily as still images. Video is an interesting research object in historical and ethnographic research. The recordings then need to be transcribed. This could be a simple interview transcription, but in disciplines like sociology, or film and media studies, it can be more complicated. In this case, the scholar would also like to annotate the source, to describe the composition of the image, the soundtrack, or the movement of the camera. The question was always: how can we watch and describe a moving image at the same time?

The moving image has a continuous linearity and makes only sense in a dynamic state. So this medium is difficult to grab and the researcher is not able to write notes and commentaries direct to it. In contrast to an image, a moving image is always bound to technical devices [1]. This fact does not make it easier to annotate them. In the digital world today moving image research is better to do as before, because we can work with only one technical device now – namely the computer. This step is comfortable, but without a corresponding software awkward to do. If the researcher works with moving images, he needs software for the video file and software for the text processing. He has to switch at least between two (proprietary) programs.

At our lab we're developing a purely web-based virtual research environment (VRE) – a system for annotation and linkage of sources in arts and humanities (SALSAH). The project originated from an art historical research project about early prints in Basel (Incunabula Basiliensia). It allows for the collaborative annotation and linking of digitized sources [2] or to define and linking special regions inside the source (e.g. region of interest on a picture).

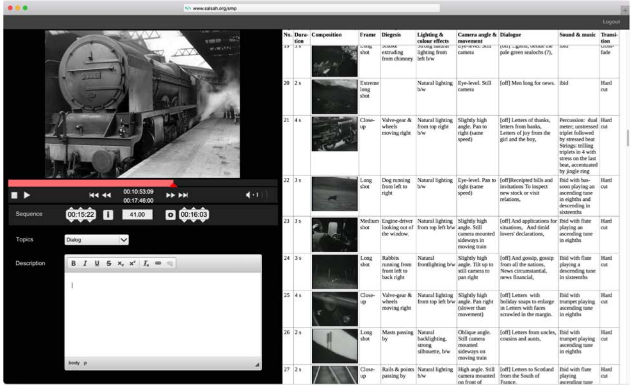
In the recent years the project grew enormously. Different projects from art history, history, media and cultural studies are using the SALSAH platform. They're all working with different kind of sources and meta data information. It doesn't matter which kind of data we're working with, because we're implemented a semantic graph database in the back-end; a triple store service based on the semantic web idea RDF (resource description framework). This data storage and long term preservation service is called KnORA: Knowledge Organization, Representation, and Annotation [3].

With the splitting of SALSAH into two parts (back-end and front-end) we have a properly RESTful API on the one hand and the possibility of various front-ends on the other side. The main page for the researchers remains SALSAH [4]. The idea is still the annotation and linkage of any kind of media. One module is the moving image transcription tool. It should be possible to bring images, text and audiovisual sources together as shown in figure 1.

An epic drama of adventure and exploration
2001: a space odyssey
2001: A SPACE ODYSSEY
A STANLEY KUBRICK PRODUCTION

Figure 1: "2001: A Space Odyssey" – Connect the movie with film posters, making of... scenes, screenplay and film stills

For a deep moving image analysis we want to have the possibility to connect a movie or just a sequence of it with related objects like screenplay, film stills or making-of...-descriptions. The connection of the movie with these additional objects and their own metadata information, enables a powerful (re)search possi- bility on the main moving image. The moving image module in SALSAH is not a standalone solution like other video analysis tools. The network behind every movie would be visible. Another difference with conventional video transcription tools is the representation of the transcription. Especially in film and media stud- ies the researcher has to describe different aspects in the movie: camera position, sound, actors, text etc. [5, 6] The result is a table based sequence protocol, as shown in figure 2 on the right hand side.



Sequence 00:15:22 41.00 00:16:03

Topics Dialog

No.	Dura-tion	Composition	Frame	Diegesis	Lighting & colour effects	Camera angle & movement	Dialogue	Sound & music	Transi-tion
19	3 s		Long shot	Smoke extruding from chimney	Strong natural lighting from left b/w	Eye-level. Still camera	[off] ...glens, beside the pale green sealochs (?),	ibid	Cross-fade
20	2 s		Extreme long shot		Natural lighting b/w	Eye-level. Still camera	[off] Men long for news.	ibid	Hard cut
21	4 s		Close-up	Valve-gear & wheels moving right	Natural lighting from top right b/w	Slightly high angle. Pan to right (same speed)	[off] Letters of thanks, letters from banks, Letters of joy from the girl and the boy,	Percussion: dual meter; unstressed triplet followed by stressed beat Strings: trilling triplets in 4 with stress on the last beat, accentuated by jingle ring	Hard cut
22	3 s		Long shot	Dog running from left to right	Natural lighting b/w	Eye-level. Pan to right (same speed)	[off]Receipted bills and invitations To inspect new stock or visit relations,	Ibid with bas-soon playing an ascending tune in eighths and descending in sixteenths	Hard cut
23	3 s		Medium shot	Engine-driver looking out of the window.	Natural lighting from top left b/w	Slightly high angle. Still camera mounted sideways in moving train	[off] And applications for situations, And timid lovers' declarations,	Ibid with flute playing an ascending tune in eighths	Hard cut
24	3 s		Long shot	Rabbits running from front left to back right	Natural frontlighting b/w	Slightly high angle. Tilt up to still camera to pan right	[off] And gossip, gossip from all the nations, News circumstantial, news financial,	Ibid with flute playing a descending tune in sixteenths	Hard cut
25	4 s		Close-up	Valve-gear & wheels moving right	Natural lighting from top left b/w	Slightly high angle. Pan right (slower than movement)	[off] Letters with holiday snaps to enlarge in Letters with faces scrawled in the margin.	Ibid with trumpet playing ascending tune in eighths	Hard cut
26	2 s		Long shot	Masts passing by	Natural backlighting, strong silhouette, b/w	Oblique angle. Still camera mounted sideways on moving train	[off] Letters from uncles, cousins and aunts,	Ibid with trumpet playing ascending tune in eighths	Hard cut
27	2 s		Close-up	Rails & points passing by	Natural lighting b/w	High angle. Still camera mounted on front of	[off] Letters to Scotland from the South of France,	Ibid with flute playing ascending tune	Hard cut

Figure 2: Example of the SALSAH movie player (top left) with the transcription tool (bottom left) and the sequence protocol (right)

The moving image is the main object in the new module. The film analyst can describe and annotate every scene with a simple transcription tool at the bottom of the SALSAH movie player (SMP). On the right hand side we can see the result of the transcription: the sequence protocol. The figure is showing just a simple example. The researcher can define the columns of the protocol by himself. Through the RDF triple store in KnORA it is possible to have the metadata depending on the research question. At the end it should be possible to export this table based sequence protocol as shown or like a subtitle file to use it with other media players.

Identifying the Same Ukiyo-e Prints from Databases in Dutch and Japanese

Taisuke Kimura
is0013hh@ed.ritsumei.ac.jp
Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

Yuting Song
gr0260ff@ed.ritsumei.ac.jp
Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

Biligsaikhan Batjargal
biligsaikhan@gmail.com
Research Organization of Science and Engineering, Ritsumeikan University, Japan

Fuminori Kimura
f-kimura@onomichi-u.ac.jp
Faculty of Economics, Management and Information Science, Onomichi City University, Japan

Akira Maeda
amaeda@media.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

Introduction

As more and more libraries, museums, galleries and archives are making their collections available online, it is becoming essential to develop methods for accessing these vast and valuable collections of cultural heritage easily and thoroughly. One of the promising approaches is to automatically identify the database records that refer to the same entity across different collections, which is called "record linkage". In the past, numerous approaches (Elmagarmid et al., 2007) have been proposed. Most of the existing approaches targeted to identify the same records in the same language. However, we aim to identify the same artworks in different languages.

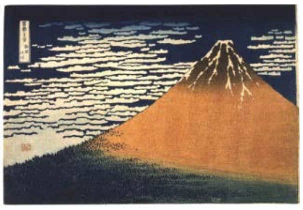
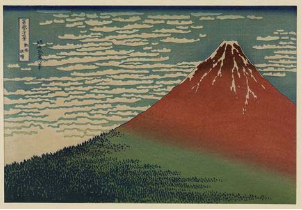
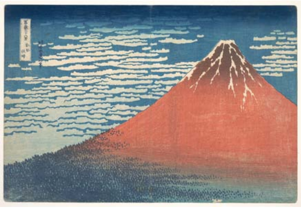
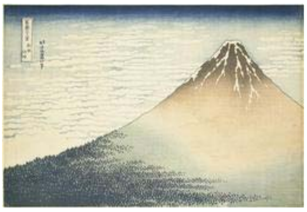
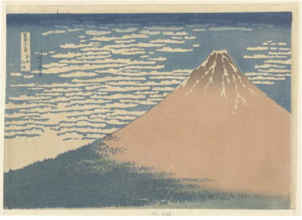
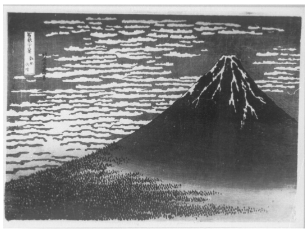
Original Ukiyo-e print	Title	Database
	凱風快晴 (original title in Japanese)	The Edo-Tokyo Museum, Japan
	Gaifū kaisei (transliteration)	The Library of Congress, USA
	South Wind, Clear Sky (in English)	The Metropolitan Museum of Art, USA
	Vent frais par matin clair (in French)	French Photo Agency, France
	Helder weer en een zuidelijke wind (in Dutch)	Rijksmuseum, Netherlands
	Fuji bei schönem Wetter von Süden gesehen (in German)	Bildarchiv Foto Marburg, Germany

Table 1. The same Ukiyo-e print in different databases with titles in different languages

In our recent work, we have developed a method for identifying the same Ukiyo-e prints from databases in English and Japanese (Batjargal et al., 2014). This method is particularly useful since Ukiyo-e, the Japanese traditional woodblock printing, is engraving and many copies or variants of one particular work were made from the same woodblock, and most of these copies were scattered around Western countries in the 19th century, and now stored in museums and galleries in these countries. Most of the metadata of these databases are available only in English or in the native language of that country. Titles are mostly written either as the transliteration of the original Japanese title, or a translation in that language. Table 1 shows an example of an Ukiyo-e print whose copies are stored in databases around the world with titles in different languages.

One of the effective approaches for identifying the same artworks from multiple image databases is to utilize image similarity calculations. Ukiyo-e.org (Resig, 2012; Resig, 2013) is the most successful example of identifying the same Ukiyo-e prints, which purely uses image similarities rather than textual data. The advantage of our approach is that we do not have to harvest all the data from the databases beforehand. This paper discusses the method for identifying the same Ukiyo-e records between Japanese and Dutch databases using textual metadata written in different languages.

Proposed approach

Here we explain our method for identifying the same Ukiyo-e records between Japanese and Dutch databases. The proposed method is divided into two main parts as shown in Figure 1. One is the literal translation of original Japanese titles into Dutch, and the other is to find the English title of the same artwork and then translate it into Dutch. The reason of having two different parts is that the translated titles of Ukiyo-e can be classified into two types, a literal translation of the original title, and a translated title, which depicts the scene or objects portrayed in the print that is not necessarily related to the original title. There are a considerable numbers of depictive titles in the translated English and Dutch titles of Ukiyo-e prints.

In the process of literal translation of original titles, first we translate the original Japanese title into Dutch by using a machine translation service (e.g. Google Translate or Microsoft Translator), and then we calculate the similarities between the literal translation and candidate Dutch titles using the similarity measure proposed in our previous approach (Kimura et al., 2015). Identified candidates are narrowed down from a Dutch database using the artist name of the original title.

In the process of using English titles, first we identify the corresponding English title(s) of the original title using the method proposed in our previous approach, then we translate the English title(s) into Dutch using a machine translation service, and then we calculate the similarities between translated Dutch title(s) and candidate Dutch titles using the same similarity measure as the literal translation process. Finally, we integrate the results of two processes and identify the same artworks that exceed a certain threshold of the similarity degree.

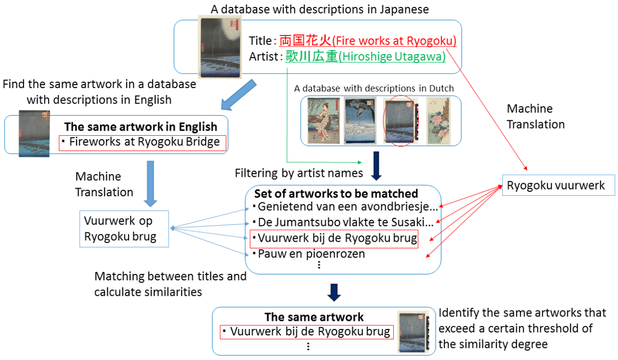

A database with descriptions in Japanese
Title : 両国花火(Fire works at Ryogoku)
Artist : 歌川広重(Hiroshige Utagawa)
Find the same artwork in a database with descriptions in English
A database with descriptions in Dutch
Machine Translation
The same artwork in English
・Fireworks at Ryogoku Bridge
Filtering by artist names
Ryogoku vuurwerk
Machine Translation
Set of artworks to be matched
・Genietend van een avondbriesje…
・De Jumantsubo vlakte te Susaki…
・Vuurwerk bij de Ryogoku brug
・Pauw en pioenrozen
Vuurwerk op Ryogoku brug
Matching between titles and calculate similarities
The same artwork
・Vuurwerk bij de Ryogoku brug
Identify the same artworks that exceed a certain threshold of the similarity degree

Figure 1. An illustration of the proposed method. Red arrows are the literal translation process and blue arrows are the process of using English titiles.

Experimental evaluation

We have conducted experiments to evaluate the proposed method. The experimental data is shown in Table 2 and the experimental results are shown in Table 3. In these experiments, we utilized the artworks of Hiroshige Utagawa.

Language	Database	Number of available Ukiyo-e prints of Hiroshige Utagawa
Japanese	The Edo-Tokyo Museum	32
English	The Metropolitan Museum of Art.	133
Dutch	Rijksmuseum	207

Table 2. Experimental data

	By employing the literal translation of original titles	By employing the English titles	Combining the literal translation and English titles
Number of correctly identified titles within top 5 ranks (percentage)	20/32 (0.6250)	14/32 (0.4375)	22/32 (0.6875)

Table 3. Experimental results

Conclusion

We proposed a method for identifying the same Ukiyo-e prints across multiple databases using textual metadata written in different languages, particularly Japanese and Dutch. By using English titles as an intermediate text, we can find not only literally translated titles but also “depictive” titles, which are common in translation of Ukiyo-e prints’ titles that cannot be identified by a simple word-to-word matching. Our preliminary experiments showed reasonable results in identifying both literally translated titles and depictive titles. As the future work, we plan to extend the proposed method to other humanities databases.

Bibliography

Batjargal, B., Kuyama, T., Kimura, F. and Maeda, A. (2014). Identifying the same records across multiple Ukiyo-e image databases using textual data in different languages, *Proceeding of the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. London, United Kingdom, pp. 193–96.

Elmagarmid, A. K., Ipeirotis, P. G. and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*. **19**: 1–16.

Resig, J. (2013). Aggregating and analyzing digitized Japanese woodblock prints. Presented at the 3rd Annual Conference of the Japanese Association for Digital Humanities, Kyoto, Japan, September 2013.

Resig, J. (2012). Japanese Woodblock Print Search **.** http://ukiyo-e.org/ (accessed 26 February 2016).

Kimura, T., Batjargal, B., Kimura, F. and Maeda, A. (2015). Finding the Same Artworks from Multiple Databases in Different Languages. *Digital Humanities 2015: Conference Abstracts*. Sydney, Australia, June 2015.

1 Million Dutch Newspaper Images available for researchers: The KBK-1M Dataset

Martijn Kleppe
martijn.kleppe@kb.nl
National Library of the Netherlands, The Hague

Elliott Desmond
d.elliot@uva.nl
University of Amsterdam

The visualisation of news through photographs has exploded since the second half of the 20th century (Kester & Kleppe 2015). Press photographs are not only being used more often in all sorts of media, historical photographs are also increasingly being reused. In some cases the reuse of these images leads to a recontextualisation of the photograph. A well-known example in the Dutch context is a photograph of Dutch socialist party leader Pieter Jelles Troelstra who is known for a failed attempt to overthrow the royal family in 1918. When Troelstra’s failed coup d’état is being discussed in history textbooks, the text is very often accompanied by a photograph of Troelstra holding a speech. However, research has shown that this photo was not made in 1918 but in 1912 when Troelstra was advocating the introduction of women’s suffrage (Kleppe 2013).

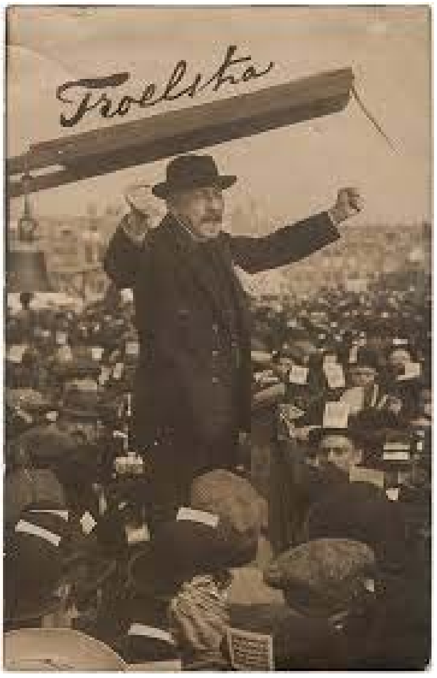


Figure 1. Photograph of Dutch socialist partyleader Pieter Jelles Troelstra giving a speech during a demonstration in 1912 while the photo nowadays often is used for Troelstra’s coupe d’etat of 1918. (photograph by Cornelis Leenheer. Source: IISG Amsterdam)

This example illustrates that in the field of history and visual culture there is a need for the thorough study of the origin and reuse of visual materials. However, methods that are employed to analyse the (re)use of visual materials are cumbersome and labour-intensive since Humanities researchers tend to analyse their sources manually (Burke 2001). To estimate the increase in the use of pressphotographs in Dutch newspapers, Kester & Kleppe (2015) e.g manually analysed a sample of 385 newspapers and 5.877 pressphotographs of the period 1870-2013, creating the *Foto’s in Nederlandse Kranten (FiNK) (Photos in Dutch Newspaper)* dataset (Kleppe, Zeijl, Kester 2014). To find the recurring use of the Troelstra image, Kleppe (2012) followed a same approach by manually analysing over 5.000 photographs in 400 history textbooks, creating the *Foto’s in Nederlandse Geschiedenisschoolboeken (FiNGS) (Photos in Dutch History textbooks)* dataset (Kleppe 2012).

While manually created and annotated datasets such as *FiNK* & *FiNGS* contain rich & well-annotated data that can be reused for all sorts of research questions, their scope remains limited given its labour-intensive creation process. To find the recurring imagery in the FiNGS dataset, Kleppe (2012) e.g. manually created and assessed all images and metadata, leading to inevitabel human errors. However, digitised historical imagery is increasingly becoming available allowing researchers to undertake the first steps in the field of 'Visual Big Data', following the footsteps of Barry Salt's study on the characteristics of opening shots of 20th century films (Salt 1974) and Scott McCloud work on the visual language of Japanese manga and comics from the West (McCloud 1994). More recent, the work of Lev Manovich on exploring large scale visual datasets such as Manga Comics (2012), Time covers, and selfies (Manovich & Tifentale 2015) is seen as a new way of what he calls doing 'cultural analytics' (Manovich 2012).

While the focus of these studies is on characteristics of the images, others using large scale image datasets focus on the recurrance of imagery in different types of contexts, aiming 1) to assess the impact of scholarly images online (Kousha 2010), 2) to analyse the reuse of digital images of cultural and heritage material on the internet (Terras & Kirton 2013) or within a closed dataset (Resig 2014; Reside 2014) and 3) to detect poetic content in historical newspapers (Lorang et al 2015).

To cater their research questions, these scholars all created visual datasets on their own. However, large datasets containing photographs that are available for researchers are scarce. Only within the Computer Vision and Natural Language Processing community we found some datasets (Ordonez et al, 2011; Chen et al, 2015a; Chen et al, 2015b; Hodosh et al., 2013), but these are mainly created for training purposes of algorithms, not for Humanities research questions.

Therefor this poster presents the *KBK-1M* dataset, that was created specifically for (Digital) Humanities researchers. This dataset contains a collection of 1,6 million captioned images of the period 1922-1994, extracted from Dutch digitised newspapers stored in the Dutch National Library (KB) newspaper archive. The images cover black & white photographs, sketches, line-drawn cartoons, and line-drawn weather forecasts and are available as year-by-year ZIP files at the National Library of the Netherlands. A request for access can be submitted to dataservices@kb.nl. On our poster, we will describe how we obtained the images and what types of research questions it could tailor.

Bibliography

Burke, P. (2001). *Eyewitnessing. The uses of images as historical evidence.* London: Cornell University Press.

Chen, J., Kuznetsova, P., Warren, D. S. and Choi, Y. (2015a). Deja image-captions: A corpus of expressive descriptions in repetition. *NAACL*, pp. 504–14.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollar, P. and Zitnick, C. L. (2015b). Microsoft COCO captions: Data collection and evaluation server. *CoRR*. abs/1504.00325.

Hodosh, M., Young, P. and Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, **47**: 853–99.

Kester, B. and Kleppe, M. (2015). Acceptatie, professionalisering en innovatie. Persfotografie in Nederland, 1837-2014. In Bardoel, J. & Wijfjes, H., *Journalistieke Cultuur in Nederland*. Amsterdam University Press, pp. 53-76.

Kleppe, M. (2013). *Canonieke Icoonfoto's. De rol van (pers)foto's in de Nederlandse geschiedschrijving*. Delft: Eburon.

Kleppe, M. (2012). *Foto's in Nederlandse Geschiedenisschoolboeken (FiNGS)*. DANS http://dx.doi.org/10.17026/dans-zfn-u8k4.

Kleppe, M., Zeijl, L. and Kester, B. (2014). *Foto's in Nederlandse kranten (FiNK).* DANS http://dx.doi.org/10.17026/dans-2cz-x7rh.

Kleppe, M. (2012). Wat is het onderwerp op een foto? De kansen en problemen bij het opzetten van een eigen fotodatabase. *Tijdschrift voor Mediageschiedenis* **14**(2): 93-107.

Kousha, K., Thelwall, M. and Rezaie, S. (2010). Can the impact of scholarly images be assessed online? An exploratory study using image identificationtechnology. *Journal of the American Society for Information Science and Technology* **61**(9): 1734-44.

Manovich, L. (2009). Cultural analytics: Visualing cultural patterns in the era of more media. *Domus* (923).

Manovich, L. (2012). How to compare one million Images? In Berry, D. M., *Understanding Digital Humanities*, pp. 249-78.

Manovich. L. and Tifentale, A. (2015). Selfiecity: Exploring Photography and Self-Fashioning in Social Media. In Berry, David M., Dieter, M. (eds), *Postdigital Aesthetics: Art, Computation and Design*, pp. 109-22.

McCLoud, S. (1994). *Understanding Comics: The Invisible Art.* New York: HarperPerennial.

Ordonez, V., Kulkarni, G. and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *NIPS*.

Reser, G. and Bauman, J. (2012). The Past, Present, and Future of Embedded Metadata for the Long-Term Maintenance of and Access to Digital Image Files. *International Journal of Digital Library Systems (IJDLS),* **3**(1): 53-64.

Reside, D. (2014). Using Computer Vision to Improve Image Metadata. *Digital Humanities 2014*.http://dharchive.org/paper/DH2014/Paper-294.xml..

Resig, J. (2013). *Using Computer Vision to Increase the Research Potential of Photo Archives*.http://ejohn.org/research/computer-vision-photo-archives/#analysis-implementations..

Salt, B. (1974). The Statistical Style Analysis of Motion Pictures. *Film Quarterly*, **28**(1): 13-22.

Terras, M. M. and Kirton, I. (2013). Where do images of art go once they go online? A Reverse Image Lookup study to assess the dissemination of digitized cultural heritage. *Selected papers from Museums and the Web North America*, pp. 237–48.

Automatic Emotion Detection for Quantitative Literary Studies: A case study based on Franz Kafka's "Das Schloss" und "Amerika"

Roman Klinger
roman.klinger@ims.uni-stuttgart.de
University of Stuttgart, Germany

Surayya Samat Suliya
suliyasa@ims.uni-stuttgart.de
University of Stuttgart, Germany

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
University of Stuttgart, Germany

Introduction

Sentiment analysis and opinion mining methods are established for automatically summarizing information shared by users in product reviews or in social media platforms like Twitter, Facebook or more specific fora (Liu 2015). These approaches can be categorized into coarse-grained and fine-grained methods: The first focus on assigning a polarity (positive, negative, neutral) and optionally an intensity to a text snippet (Täckström and McDonald 2011; Pang and Lee 2004). The latter additionally aim at detecting the opinion holder (for instance a specific person mentioned in a news article) and the target (for instance a specific aspect of a product in a review) (Hu and Liu 2004; Popescu and Etzioni 2005; Jakob and Gurevych 2010).

Transferring such methods to the analysis of literature leads to at least two questions: Firstly, are polarities for this domain as helpful as for the analysis of reviews? Secondly, how can such methods from sentiment analysis be improved, and what can they contribute to literature analysis?

Regarding the first aspect, resources to measure the occurrence of words which are associated with different emotions have been developed for English but, to the best of our knowledge, not for German (Mohammad et al. 2015). Secondly, it should be noted that research in German sentiment analysis is still comparably limited (counter examples are Ruppenhofer et al. 2014; Klinger and Cimiano 2015; Remus, Quasthoff, and Heyer 2010). In addition, sentiment analysis has mainly focused on the Web, like social media, and product reviews. However, the analysis of emotions and sentiment in literature has been proven to be of interest and value (Mellmann 2007; Winko 2003). A prerequisite for a quantitative approach is that emotions are (at least to some extend) a surface phenomenon (Hillebrandt 2011, p. 154), i.e., that words carry information such that it is possible to infer "private states" of specific emotions (Wiebe, Wilson, and Cardie 2005).

Our two main contributions are: (a) We make German dictionaries of words associated with seven fundamental emotions publicly available, and (b) perform a case study on Kafka's "Amerika" and "Das Schloss" regarding emotion analysis to support literature studies with a focus on complex non-factual phenomena and the analysis of personality traits. All resources and software used in this paper are made publicly available at http://www.roman-klinger.de/emotion/.

Materials and Methods

The goal of this work is to detect different emotions represented in literary texts. Psychological research offers different models to categorize emotions. The most common ones include Plutchik's wheel of emotions (Plutchik 2001) and Ekman's definition of fundamental emotions (Ekman 1999). A discussion of relevant context is offered by Russell (Russell 1991). We opt for roughly following the structure of Ekman's definition of emotions and focus on *anger* (Wut), *disgust* (Ekel), *fear* (Angst), *enjoyment* (Glück), *sadness* (Trauer), and *surprise* (Überraschung) and *contempt* (Verachtung).

To track emotions over the whole text, we assign an emotion score es(*e*, **t***ab*) to a subset of consecutive tokens **t***ab* from textual position *a* to position *b* as

$$\mathrm{es}(e, \mathbf{t}_{ab}) = \frac{1}{|D_e|} \sum_{i=a}^{b} \mathbf{1}_{t_i \in D_e}$$

where *De* is a dictionary containing words expressing the specific emotion *e* and 1 t ∈ *D* is 1 if and only if $t_i \in D_e$ and 0 otherwise. This score corresponds to the number of tokens which are in a window and in the respective emotion dictionary, normalized by the dictionary size.

To track the development of the emotions over the whole text, we apply a sliding window approach which is parameterized by window size w such that *b* = *a* + *w* – *1* (which can be interpreted as a smoothing parameter). To allow for a character oriented analysis, we assign an emotion score as in the sliding window, but for windows around each mention of such character in the text, with an additional normalization based on number of character mentions. Each token and dictionary entry is normalized by mapping to lower-case and stemming with the Porter stemmer (Porter 2001).

As a resource for the emotion dictionaries, two authors of this paper manually selected words from different sentiment polarity, subjectivity, and emotion resources in German and English (translated to German) into the emotion categories (Waltinger 2010a; Waltinger 2010b;

Remus, Quasthoff, and Heyer 2010; Mohammad and Turney 2013). We semi-automatically enriched this resource with synonyms (Naber 2005; Wermke, Kunkel-Razum, and Scholze-Stubenrecht 2010).

Experiments and Conclusions

As an estimate for the difficulty of emotion assignments, we performed an annotation experiment of 300 words (stratified sample from all emotions in the dictionary mentioned above) with fluent speakers of German. In 85 % of all words two out of three annotators agree on the same emotion, however, only in 46 % of of all words, three annotators agree on the associated emotion.

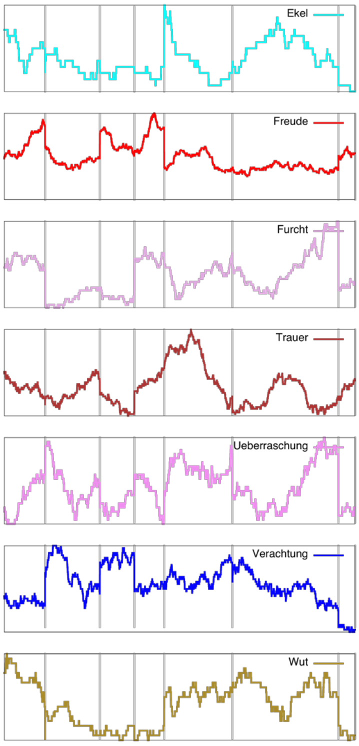


Figure 1: Tracking fundamental emotions in Kafka's "Amerika". The vertical grey lines indicate the start of each chapter. The horizontal axis denotes sequential positions in the text. The vertical axis denotes relative frequencies of the respective emotion.

As a use-case, we apply our methods to Franz Kafka's "Der Verschollene" ("Amerika") and "Das Schloß". Especially the latter is interesting as a comprehensive emotion-focused manual analysis is available (Hillebrandt 2011). It is narrated in third person and interesting from an emotion analysis point of view, as attribution of specific emotions to the protagonist is difficult (Hillebrandt 2011, p. 165).

The development of emotions in Figures 1 and 2 visualize the outcomes of our analyzes. In "Das Schloss", the strong increase of surprise towards the end is striking (most indicative words are "neu", "schnell", "plötzlich", "ungeduldig"). Another example for an eye-catching peak of fear is shortly after start of chapter 3 ("ängstlich", "Gefahr", "unruhig", "Gewalt"). In "Amerika", one striking characteristic is the decrease of enjoyment after a peak in chapter 4 ("gut", "Mutter", "glücklich") followed by disgust in chapter 5 ("unerträglich", "Elend", "schrecklich", "beschmutzt"). Emotions for each mention of a selection of characters in "Amerika" and "Das Schloss" are shown in Figures 3 and 4.

Acknowledgements

This project has been partially funded by the project CRETA (Zentrum für reflektierte Textanalyse, http://www.creta.uni-stuttgart.de/), funded by the German Ministry of Education and Research.

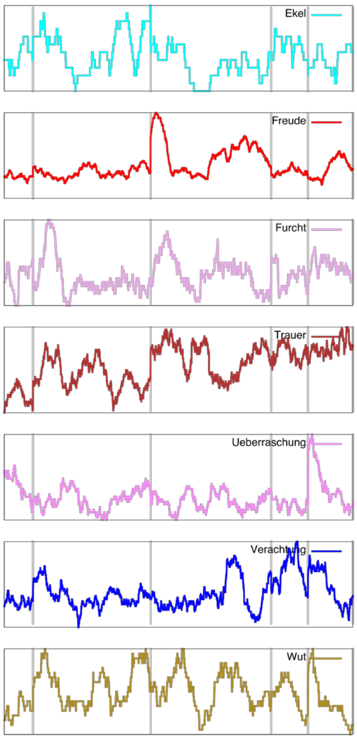


Figure 2: Tracking fundamental emotions in Kafka's "Das Schloss". The vertical grey lines indicate the start of chapters 3, 13, 16, and 19, respectively. The horizontal axis denotes sequential positions in the text. The vertical axis denotes relative frequencies of the respective emotion.

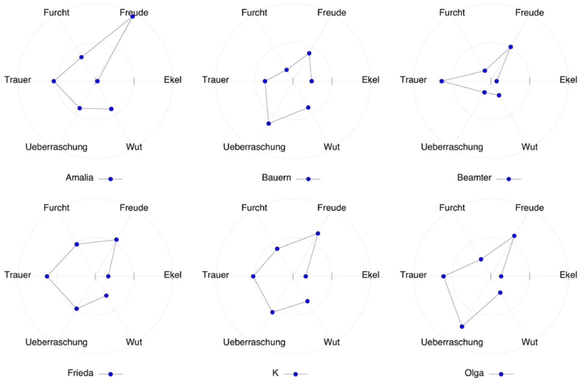


Figure 3: Character Profiles for "Das Schloss". ("Verachtung" has been omitted, as it can be seen as a mixture of the other emotions.)

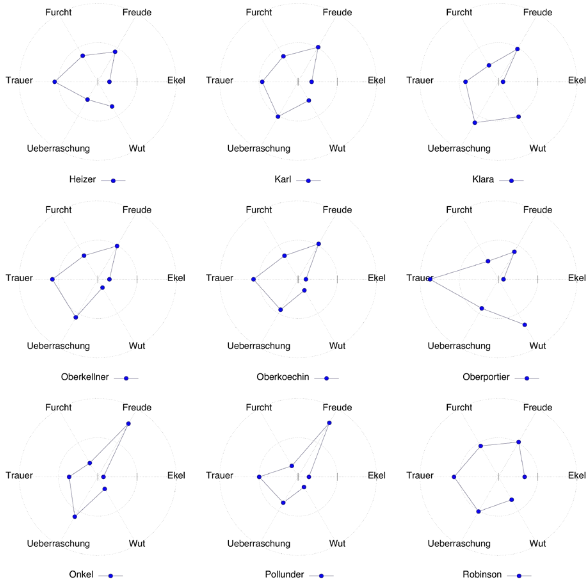


Figure 4: Character Profiles for "Amerika". ("Verachtung" has been omitted, as it can be seen as a mixture of the other emotions.)

Bibliography

Ekman, P. (1999). "Basic Emotions". In: *Handbook of Cognition and Emotion*. Ed. by M Dalgleish T; Power. Sussex, UK: John Wiley & Sons.

Hillebrandt, C. (2011). *Das emotionale Wirkungspotenzial von Erzähltexten*. Deutsche Literatur - Studien und Quellen. Berlin, Germany: Akademie Verlag.

Hu, M. and Liu, B. (2004). "Mining and summarizing customer reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. Seattle, WA, USA: ACM, pp. 168–77.

Jakob, N. and Gurevych, I. (2010). "Extracting opinion targets in a single- and cross-domain setting with conditional random fields". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 1035–45.

Klinger, R. and Cimiano, P. (2015). "Instance Selection Improves Cross-Lingual Model Training for Fine-Grained Sentiment Analysis". In: *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China: Association for Computational Linguistics, pp. 153–63.

Liu, B. (2015). *Sentiment Analysis*. Cambridge University Press.

Mellmann, K. (2007). *Emotionalisierung – Von der Nebenstundenpoesie zum Buch als Freund*. Vol. 4. Poetogenesis - Studien zur empirischen Anthropologie der Literatur. Münster, Germany: Mentis Verlag.

Mohammad, S. M. and Turney, P. D. (2013). "Crowdsourcing a Word-Emotion Association Lexicon". In: *Computational Intelligence*, **29**(3): 436–65.

Mohammad, S. M., Zhu, X., Kiritchenko, S. and Martin, J. (2015). "Sentiment, emotion, purpose, and style in electoral tweets". In: *Information Processing & Management*, **51**(4): 480–99.

Naber, D. (2005). *OpenThesaurus: ein offenes deutsches Wortnetz*. http://danielnaber.de/ publications/gldv-openthesaurus.pdf (visited on 02/17/2015).

Pang, B. and Lee, L. (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". In: *Proceedings of the 42nd Meeting of the Association for Computational Linguistics, Main Volume*. Barcelona, Spain, pp. 271–78.

Plutchik, R. (2001). "The Nature of Emotions". In: *American Scientist*, **89**: 344–50.

Popescu, A.-M. and Etzioni, O. (2005). "Extracting Product Features and Opinions from Reviews". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Associ- ation for Computational Linguistics, pp. 339–46.

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. http://snowball. tartarus.org/texts/introduction.html.

Remus, R., Quasthoff, U. and Heyer, G. (2010). "SentiWS – a Publicly Available German- language Resource for Sentiment Analysis". In: *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pp. 1168–71.

Ruppenhofer, J., Klinger, R., Struß, J. M., Sonntag, J. and Wiegand, M. (2014). "IGGSA Shared Tasks on German Sentiment Analysis". In *: Workshop Proceedings of the 12th Edition of the KONVENS Conference*. Ed. by G. Faaß and J. Ruppenhofer. Hildesheim, Germany: Uni- versity of Hildesheim.

Russell, J. A. (1991). "In Defense of a Prototype Approach to Emotion Concepts". In: *Journal of Personality and Social Psychology*, **60**(1): 37–47.

Täckström, O. and McDonald, R. (2011). "Semi-supervised latent variable models for sentence- level sentiment analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 569–74.

Waltinger, U. (2010a). "GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta.

Waltinger, U. (2010b). "Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identi- fication Combining Machine Learning And Subjectivity Features". In: *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10)*. Valencia, Spain.

Wermke, M., Kunkel-Razum, K. and Scholze-Stubenrecht, W., (eds). (2010). *Duden – Das Synonymwörterbuch*. Mannheim, Zürich: Dudenverlag.

Wiebe, J., Wilson, T. and Cardie, C. (2005). "Annotating Expressions of Opinions and Emotions in Language". In: *Language Resources and Evaluation*, **39**(2): 165–210.

Winko, S. (2003). *Kodierte Gefühle: Zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900*. Erich Schmidt Verlag.

Creating A Management Plan For A Cultural Heritage Project – Best Practise

Carina Koch
carina.koch@uni-graz.at
University of Graz, Austria

The project "Repository of Styrian Cultural Heritage" aims to build a digital archive of cultural heritage objects to make digitized and scholarly annotated collections from various institutions (archives, museums or universities) available for the public. To implement accurate retrieval functionalities and special visualizations a consistent data pool is indispensable. Furthermore, to guarantee data longevity and the possibility to analyse, re-use or add to the collected data, a well-documented management plan is of great importance. Contrary to possible assumptions that the increasing number of similar portals means that comparable documentations already exist in abundance, most portals either do not have any kind of documentation (e.g. Kulturerbeportal-Niedersachsen, BLO, DG-Kulturerbe) or strongly focus on details (e.g. metadata and mapping guidelines or specify controlled lists for comparability) (e.g. bavarikon, DDB, Europeana, museum-digital, MIMO). They do not offer extensive guidelines, especially for the creation of digital collections from scratch. At the Centre for Information Modelling in Graz we designed a management plan for this project according to the OAIS reference model and taking into account aspects concerning research data management (e.g. Puhl et al., 2015), for all partners involved.

Because we gather data from more than one preservation infrastructure (e.g. FEDORA-based repository GAMS) and deal with a variety of resources, from text-centred materials and images to museum objects and artefacts from various contexts (e.g. criminology or archaeology), the designed strategy is generic and expandable. The poster will show systematic strategies regarding different aspects ranging from evaluation to content curation and data re-use. The steps include:

Analysis & Evaluation: The starting-point is to evaluate and analyse the structure of the diverse sources. It involves a discussion about the goal and scope of the project in general as well as possible publication scenarios and functionalities for the collected data. At this stage, communication between technically skilled humanists and specialists for the respective collections is crucial. This collaboration saves time and effort in later stages of the project. Based on this foundation, workflows and data models exactly suited to the needs of the specific institution, including legal aspects, can be established.

Modelling & Metadata Generation: After these evaluations, a data model, including obligatory object descriptions as a minimum requirement for all object types along with rules for metadata creation, has to be developed. Furthermore the depth of the annotation and the use of controlled vocabularies for semantic enrichment have to be defined.

Quality Control & Preserving Data: Quality control of data is an integral part and takes place at various stages, during data collection or digitization. Already existing digital data has to be reviewed and if necessary revised according to the defined rulesets. In this respect, the documentation of data provenance (origins of data, revision agreements and transformation scenarios) is an essential step. Gathered and harvested data has to be transformed from proprietary file formats to suitable storage formats. Internationally recognized standards like DC for basic descriptive metadata, TEI for manuscripts or LIDO for museum objects are recommended to maximize possible re-use and interoperability of the data. For the web portal, all objects are going to be mapped to EDM. The model forms the common ground for the general object description and semantic enrichment.

For long-term preservation, platform-independent systems and open-source software like a Fedora-based repository should be used (e.g. Stigler, Steiner 2015). All objects need persistent identification, which ensures its availability and citability.

Access & Dissemination: Based on consistent data, various search options (filter, facet or advanced) can be implemented. Next to a general representation for all artefacts, elaborate object-specific functionalities and visualizations, like the possibility to thumb through manuscripts or view postal routes of correspondence on a historic map, can be offered.

Adding & Re-use: The possible re-use of data and the addition of new objects or scientific findings as well as the further enrichment of metadata also have to be considered. Adding information according to the defined guidelines guarantees the same quality for re-use, whether for follow-up research, teaching or browsing.

The poster will generally introduce the approaches to a common web portal for a variety of digital resources. It will present a best practice model for an interdisciplinary, cross-institutional cultural heritage project, based on experiences concerning data preservation, consistent metadata description and enrichment. In this respect it also provides guidance to implement a long-lasting expandable digital archive.

Bibliography

Bavarikon. http://www.bavarikon.de/object/bav:BSB-CMS-0000000000000605, (accessed 30 October 2015).

BLO – Bayerische Landesbibliothek Online. http://www.bayerische-landesbibliothek-online.de, (accessed 30 October 2015).

CCSDS – Consultative Committee for Space Data Systems (2012). OAIS-MAGENTA BOOK (2012). http://public.ccsds.org/publications/archive/650xom2.pdf, (accessed 30 October 2015).

DC – Dublin Core. http://dublincore.org/specifications, (accessed 30 October 2015).

DDB – Deutsche Digitale Bibliothek. https://api.deutsche-digitale-bibliothek.de/doku/display/ADD, (accessed 30 October 2015).

DG-Kulturerbe. http://www.dgkulturerbe.be, (accessed 30 October 2015).

EDM – Europeana Data Model. http://pro.europeana.eu/page/edm-documentation, (accessed 30 October 2015).

Europeana: http://pro.europeana.eu/, (accessed 30 October 2015).

Kulturerbeportal-Niedersachsen. http://kulturerbe.niedersachsen.de/viewer/start/, (accessed 30 October 2015).

LIDO – Lightweight Information Descriping Objects. http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf, (accessed 30 October 2015).

MIMO – Musical Instrument Museums Online. http://www.mimo-international.com/MIMO/how-to-join.aspx, (accessed 30 October 2015).

Museum-digital: http://www.museum-digital.de/index.php?t=einmaleins, (accessed 30 October 2015).

Puhl, J. et al. (2015). Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften, PURL. http://resolver.sub.uni-goettingen.de/purl/?dariah-2015-4, (accessed 30 October 2015).

Repository of Styrian Cultural Heritage. http://wissenschaftserbe.uni-graz.at/en, (accessed 30 October 2015).

Stigler J. and Steiner E. (2015). GAMS and Cirilo Client Policies, documentation and tutorial. http://gams.uni-graz.at/doku, (accessed 30 October 2015).

Stigler J. Cirilo Client. https://github.com/acdh/cirilo, (accessed 30 October 2015).

TEI – Text Encoding Initiative. http://www.tei-c.org, (accessed 30 October 2015).

Geography Of Russian Poetry: Countries And Cities Inside The Poetic World

Elizaveta Kuzmenko
lizaku77@gmail.com
National Research University Higher School of Economics, Russian Federation

Boris Orekhov
nevmenandr@gmail.com
National Research University Higher School of Economics, Russian Federation

Our paper is dedicated to two major problems: the first problem is the digital one and the second problem is connected to the humanities. The first problem involves automatic extraction of named entities, and the second problem is connected to the usage of toponyms in poetic texts. Correspondingly, our research comprises two parts: automatic processing of a huge amount of texts from the corpus of Russian poetry and revealing major trends in the functioning of toponyms during the history of the Russian poetry from 18 to 20 centuries.

Our research is based on the data from the poetic corpus which is a part of the Russian National Corpus[1]. This corpus includes the main texts belonging to the Russian poetry from all the periods of its history, up to the 20th century. The principles of text selection in the poetic corpus are described by its creators (Grishina et al., 2009). The size of the corpus is approximately 11 million word tokens.

Up to the present moment, research papers considering toponyms in the Russian poetry described a concrete toponym from the perspective of an isolated text or a particular author (see, for example, Mednis, 1999). Our approach is quite different: we describe tthe geography of Russian poetry as a whole, consistently to the framework of distant reading (Moretti, 2005; Moretti, 2013). Thus, the result demonstrates global trends in the usage of toponyms in Russian poetry as a system.

We used two different technologies to extract geographic entities from poetic texts, and the comparison of these two approaches is one of the results of our research. The first technology is a proprietary commercial software Textocat[2], which is based on machine learning with the use of nonfictional texts as a training sample. The creators of this software claim that the F1-measure for the retrieval of named entities is 0.75. However, it is expected that the performance would be much lower in the case of poetic texts, because the language of poetry differs radically from the language of prose.

The second approach we use is a self-made tool for the extraction of toponyms based on the dictionary of geographical names. We are forced to create such a tool because there is no open-source software for the extraction of toponyms for Russian. As a basis for our dictionary of geographical names, we use the list of toponyms from the Wikipedia.We compared the figures retrieved with our approach to the ones resulting from Textocat. We used for evaluation a sample of toponyms consisting of countries and cities. The comparison showed that Textocat retrieves only 25.7% of country names and 19.3% of city names that are found with our tool. In addition, Textocat makes a lot of mistakes; for example, locative pronouns *там* 'there' and *где* 'where' are retrieved among geographical entities. The words *страна* 'country' and *город* 'city' are also included by Textocat in the list of found toponyms.

As we can see, the dictionary-based approach proves to be more efficient, and in further results we consider only the data extracted with this method.

First, we will look in detail on the names of countries extracted from poetic texts. The distribution of mentioning countries is presented in Table 1 (six most popular countries are taken):

Country	Number of times it is mentioned
Russia	2744
France	283
Italy	241
Poland	201
Lithuania	160
Greece	159
Egypt	151

Table 1. The most frequently mentioned countries

The distribution of names for all the world countries is presented on Figure 1. Before building this map, the frequencies of countries' names were normalized, so we counted the number of poetic pieces in which the name was met, not the occurrences.

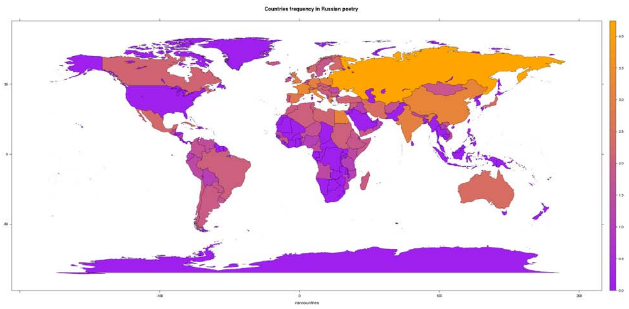


Figure 1. Frequency of countries' names in the Russian poetry

It is not surprising that Russia takes the first place on this list. The top of the list is occupied mainly by the European countries. The second place goes to France, because its influence on Russian culture was immense: in the 19th century French was even the main language of communication for Russian aristocracy. Italy can be found on the third place, despite the fact that it is very important for the poetic mythology in the 19th century, and it was the main geographical location for the poetry of the eminent Russian poets Batyushkov and Baratynsky.

It should also be mentioned that the Russian poetry does not favor exotic countries and it is primarily occupied with the European neighbors of Russia (Poland, Lithuania, Greece). The only African country in the top of «poetic» countries is Egypt, which is renowned for its ancient mythology and pyramids with considerable poetic potential.

The second exotic country in our list is India, and it is followed by countries of the specific «Russian East», which includes Georgia and Iran. The most popular country from the Middle East is Lebanon. If we take a look at the contexts in which Lebanon is used, we see that this country is mostly mentioned due to the cedars of Lebanon. It is unexpected that Lebanon appeared to be more frequently mentioned than the Northern European countries (Finland, Norway, Denmark).

Now we will consider mentionings of city names in Russian poetry. The frequency of names for European cities can be seen on Figure 1. This map reflects mentioning of cities with frequency lower than 690. Thus, we drop such cities as Moscow (with frequency of 2470), Rome (with frequency of 1135), Paris (771), and Saint Petersburg (695). Let us note that Rome is more frequent in Russian poetry than Paris, although France dominates Italy in the list of countries' mentionings. Also, we do not mark on the map those cities whose frequncy is lower than 4.

As we can see from the visualization, the most «poetic» cities from the point of view of Russian poets are concentrated near Moscow and Saint Petersburg, and also in Ukraine and Northern Italy. Ukraine was a part of Russia during the history of Russian poetry, but the specifically Russian territory demonstrates uneven distribution of mentionings, whereas Ukraine is densely populated with poetic cities. The Crimea draws attention as the most «charged» with poetic cities, though it is not a big territory itself. Sea coasts of Southern Europe generally provide us with cities popular among Russian poets.

Continental Europe is not frequently mentioned in the poetry, with an exception of the capital cities of Prague and Warsaw. The blank spaces can be found throughout the territories of France and Germany. Scandinavian cities also don't have considerable amount of mentionings within the history of Russian poetry.

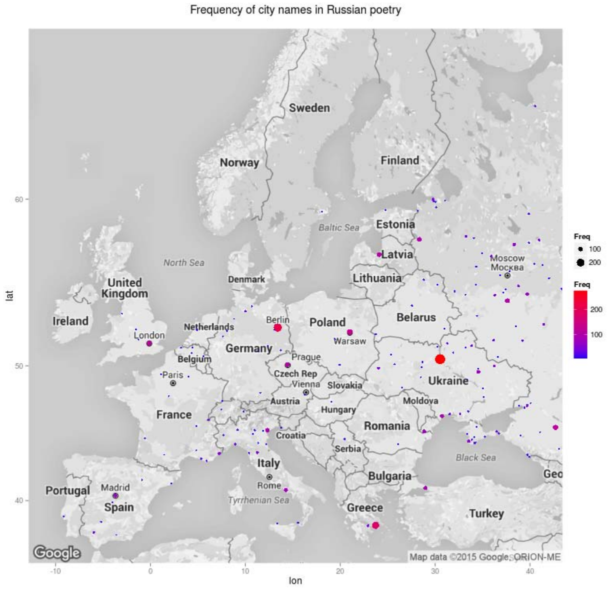


Figure 2. Frequency of city names in the Russian poetry

Another interesting opposition lies in the distribution of mentionings for Romanic and Germanic cities. As we can see, Russian poets prefer the cities of Italy, France, Spain and Belgium, whereas cities of Britain, Germany and Netherlands appear to be less poetic. Probably, the reason underlying this distribution is not the language, but the confession. Apparently, Russian poets prefer catholic countries to the protestant ones, and Vatican itself can be found on the 17th place judging by the frequency of mentioning countries in the Russian poetry.

To sum up, mentioning of toponyms presents interesting data which can be used to reveal trends in the poetry, and those trends help to describe Russian poetry as a system. It is impossible to notice such patterns with manual or visual analysis of poetic texts, but it can be achieved through digitalization of poetry and computational analysis of the texts.

In future we plan to extend our research towards extraction and interpretation of other types of geographic entities: water bodies, geographical places, regions, streets, eminent buildings and locations. This would require dealing with the change of names throughout the time periods. We also plan to investigate the usage of geography in the particular poets' works.

Bibliography

Grishina, E., Korchagin, K., Plungian, V. and Sitchinava, D. (2009). *Poeticheskij korpus v ramkah NKRYa: obschaja struktura i perspektivy ispol'zovanija. 'Natsional'nyj korpus russkogo jazyka: 2006-2008. Novye rezul'taty I perspektivy'.* Saint Petersburg, pp. 71-113. [Poetic Corpus in RNC: general structure and using perspectives]

Mednis, N. (1999). *Venecija v russkoj literature.* Novosibirsk. [Venice in the Russian Literature].

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract models for a literary history.* Verso.

Moretti, F. (2013). *Distant Reading.* Verso.

Notes

[1] http://ruscorpora.ru/search-poetic.html

[2] http://textocat.ru/

The Holy and the Godless – Cultural Stereotypes Featured in the Language of the Polish Medieval Hagiography. A Corpus-based Study

Anna Ledzińska
ledzinsk.a@gmail.com
Institute of Polish Language, Polish Academy of Sciences

My poster presents a work-in-progress focused on the language of the Polish medieval Latin hagiography. The main subject of my research are the words and expressions denoting the holy and the godless extracted from the corpus. Therefore my poster concentrates on four main topics:

1. The Corpus of the Polish Medieval Hagiography itself

2. Delimitation of the field of study within the linguistic material in relation to historical circumstances and cultural landscape of the Polish Middle Ages

3. Methodology of the research

4. Questions arising and first results.

1. At present the Corpus consists of three main types of texts – *Vitae, Miracula, Translationes - Lives, Miracles and Translations of the bodies of the Saints*. It comprises about thirty texts, it is nearly 500 000 words. Firstly, I would like to explain the reasons of such a selection of the components resulting in the expected homogeneity of the whole body of texts (from the linguistic as well as cultural point of view): **a.** established time frame and limitations concerning authorship and place where a given text was written (Starnawski, 1993), **b.** the phenomenon of mixing of the literary genres within particular groups of texts, **c.** last but not least - the preservation of the texts and the existence of critical editions of the majority of them. Secondly, I describe technical parameters of the Corpus (Piotrowski, 2012): XML format, microstructure and typography of the texts encoded according to the TEI Guidelines (P5), morphosyntactic mark-up with TreeTagger (Schmid, 1994). The Corpus is being analyzed using multi-tool platform TXM (Heiden., Magué, Pincemin, 2010).

2. The starting point of the investigation is the scheme of the seven virtues and seven vices with their subdivisions, popular in the European Middle Ages, sometimes represented in manuscripts in the form of two seven-branched, multi-leaved trees, each part of which is meticulously subscribed with a respective Latin expressions (Marchese, 2013). This traditional set of vices and virtues is being compared with the list of words of the Corpus. At the same time, apart from the historical and literary data mentioned above, other elements of character description, such as epithets, comparisons and attributes connected strictly with each of the saints and their opponents are searched and analyzed (Sinclair, 2003; Stubbs, 2001). As a result I expect to obtain a huge lexical resource containing the most eminent vocabulary of the subject, which will open the prospect of further research. It should be noted that the antithetic tension between the good and the bad one was particularly emphasized in the tradition about the Bishop Stanislaw murdered by the King Boleslaw II the Generous. Because of the great importance of the Saint himself and of the wide spread narration about the Episcopal/royal conflict, it is very interesting to observe how it is echoed (or ignored) in the histories of other saints.

3. Since the texts considered were written in a very long period - between the end of the 10th and the end of the15th century - the linguistic material is studied mainly in a diachronic perspective, although there are other possible applications, which require dividing the Corpus into parallel sub-corpora. The possible research questions in this case include: **a.** comparing and contrasting the image of male and female Saints, **b.** comparing Dominican, Franciscan and other traditions, **c.** showing the dialogue between the hagiographers of the two Patrons of the Polish Kingdom: St. Stanislaw and St. Adalbert, **d.** shedding some more light on Polish saints and sinners in the context of the European Middle Ages (in the future).

4. The main question is two-fold: what are the most frequent features and virtues ascribed to the Polish medieval saintly men and women and, on the contrary - features and vices used to describe sinners/bad people? Then, does it change in time and is this change congruent with what we know about the transformation of the models of sanctity and the image of "the Other"? All these problems, as well as other issues mentioned above (point 3) can only be solved by a complex analysis of textual data combining methods of the corpus linguistics and those of the literary and historic studies. The Corpus I present is meant to be a starting point for such complex studies.

Bibliography

Gașpar, C., Miladinov, M. and Wood, I. (2013). *Saints of the Christianization Age of Central Europe: Tenth to Eleventh Centuries.* Central European University Press.

Heiden, S., Magué, J.-P. and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *The 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*. Edizioni Universitarie di Lettere Economia Diritto, pp. 1021–32.

Marchese, F. T. (2013). Virtues and Vices: Examples of Medieval Knowledge Visualization. *Proceedings of the 17th International Conference on Information Visualization: IV'13.* (London, UK, July 16-18, 2013), IEEE Computer Society. Los Alamitos, CA, pp. 359-65.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction.* Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book.* Taylor & Francis.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts.* Morgan & Claypool Publishers.

Plezia, M. (1987). *Średniowieczne żywoty i cuda patronów Polski*. Instytut Wydawniczy Pax.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, pp. 44–49.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford University Press.

Sinclair, J. (2003). *Reading concordances : an introduction*. New York: Pearson/Longman.

Starnawski, J. (1993). *Drogi rozwojowe hagiografii polskiej i łacińskiej w wiekach średnich*. Kraków: Polskie Towarzystwo Teologiczne.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Wiley-Blackwell.

Witkowska, A. (1999). *Nasi święci: polski słownik hagiograficzny*. Księgarnia Św. Wojciecha.

Taking VIVO into the Past: Adapting the VIVO Researcher Profile System to Historical Persons

Thea Lindquist
thea.lindquist@colorado.edu
University of Colorado Boulder, United States of America

Alex Viggio
alex.viggio@colorado.edu
University of Colorado Boulder, United States of America

This poster will explain the concept behind and report preliminary results of an in-progress project, VIVO for Historical Persons (VIVO4HP). This project is an experiment to reuse and extend the VIVO-Integrated Semantic Framework (ISF) ontology to accommodate historical persons, using early Stuart diplomats (1603-1649) as a use case. VIVO4HP investigates whether VIVO can be reasonably adapted to represent and facilitate discovery of structured biographical data about historical persons. It represents a foray into the adaptation of existing systems to historical purposes interesting to humanists, the VIVO user community, and semantic web researchers.

VIVO (Börner et al., 2012) is an open-source semantic application that represents modern academic research communities. VIVO's purpose as a researcher profile system is to represent scholars, their interests, activities, and accomplishments, and the networks among them. The VIVO-ISF ontology interlinks researcher biographical data in meaningful ways to enable discovery, analysis, and visualization of scholarly networks. It does this, in part, by reusing and extending a number of established ontologies already integrated with the VIVO-ISF, such as Friend of a Friend (FOAF) and the Bibliographic Ontology (BIBO). Additionally, we are evaluating external ontologies relevant to the use case that could be imported into the system. Given VIVO's origin in academia and active user community, a number of efforts are underway to extend the ontology in areas like aerospace and agriculture; however, little experimentation has occurred in the humanities let alone in the historical context.

Our investigation aims to assess whether VIVO can be reasonably adapted to represent and facilitate discovery of biographical data about historical persons, a humanities use case that corresponds closely to its original purpose. Our initial work focuses on historical persons belonging to a specific professional community – early Stuart diplomats, that is, the diplomats who served James I and Charles I of England in the first half of the seventeenth century. The test data is derived from the *Oxford Dictionary of National Biography* (*ODNB*) (Matthew and Harrison, 2004; Goldman, 2005-13), the standard British biographical source. They are supplemented by event data on the diplomats' missions from *A Handlist of British Diplomatic Representatives* (Bell, 1990). VIVO4HP will make the data available as linked data, a flexible format that encourages data sharing and integration. Thus the data will lend themselves to a variety of projects and applications for subsequent analysis, evaluation, and visualization.

We are undertaking this work as a multi-stage, iterative process. The first step is to manually create profiles in VIVO for a limited number of diplomats to identify issues with the source data, default ontology, data mapping and transformation, and online display. The second is to make adjustments to the ontology and online display to address gaps. The third is to automatically ingest *ODNB* data into VIVO, using custom scripts where possible to address data mapping and transformation issues. The final step is to augment the profiles with other data sources, such as historical sources, other web sites, and linked data.

A variety of issues surface related to working with humanities data (Posner, 2015) in an environment that was not necessarily built to accommodate them. General issues with historical data include dealing with data ambiguity (e.g., dates, spelling variations), incomplete biographical information, individuals' identities (e.g., using noble titles instead of their personal names), and historical geographies. The *ODNB* data was created with another purpose in mind and thus can be incomplete and either untagged or not tagged to the specificity necessary for certain types of data that might be desirable to represent for our use. Finally, since VIVO is meant to represent a researcher's professional life, it does not incorporate some of the personal, political, or social aspects desirable for representing historical persons.

How we are able to deal with these issues, and the level of intervention required, will play an important role in our assessment of VIVO's utility for the discovery and representation of historical persons and whether the outcome merits the investment required to further extend it for this purpose. We hope that VIVO4HP will provide a basis that others can build upon to represent, facilitate discovery of, and share data about historical persons in the online environment.

Bibliography

Bell, G. M. (1990). *A Handlist of British Diplomatic Representatives, 1509-1688*. Royal Historical Society Guides and Handbooks, 16. London: Royal Historical Society.

Börner, K., et al. (2012). *VIVO: A Semantic Approach to Scholarly Networking and Discovery*. Synthesis Lectures on the Semantic Web: Theory and Technology. [San Rafael, Calif.]: Morgan & Claypool.

Goldman, L. (ed). (2005-13). *Oxford Dictionary of National Biography*. Online edn. Oxford: Oxford University Press. http://www.oxforddnb.com/. Accessed 28 February 2016.

Matthew, H. C. G. and Harrison, B. (eds). (2004). *Oxford Dictionary of National Biography*. Oxford: Oxford University Press.

Posner, M. (2015). Humanities data: a necessary contradiction. *Miriam Posner's Blog*. 1 November. http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/. Accessed 28 February 2016.

Notes

[1] http://demo.vivo4hp.org:8080/

[2] https://wiki.duraspace.org/display/VIVO/VIVO-ISF+Ontology

[3] http://www.vivoweb.org/

Digital Humanities Pedagogy as Digital Liberal Arts: A Framework for Curriculum Development

Brandon Locke
blocke@msu.edu
Michigan State University, United States of America

The Lab for the Education and Advancement in Digital Research (LEADR) is a new initiative of the Michigan State University Department of History, Department of Anthropology, and Matrix. LEADR's role is to integrate digital humanities and digital social science methods into courses in affiliated departments. In its first two years of operation, LEADR has partnered with 40 instructors to add digital components to 63 courses, interacting with around 1600 students. These partnerships varied from a single hour-long workshop to semester-long experiential courses in digital research.

While some courses are specifically designed as Digital History or Digital Anthropology, most of these courses are standard Anthropology and History courses composed of majors from all across campus. For courses such as these, LEADR has begun to frame these digital projects and activities as an extension of the core values of the liberal arts: the skills essential for individuals to participate in civic life, including the ability to seek information, to engage, analyze, and criticize this information, and to communicate their views in an effective manner. It is these fundamental liberal arts skills, in the context of 21st century information and communication that LEADR develops through on to disciplinary, content-based coursework.

This poster will illustrate LEADR's framework for developing Digital Humanities and Social Science curricula that contribute to understandings of emerging disciplinary methods while developing essential skills for students in broad-ranging domains. The four competency goals are:

Information Literacy is defined by the American Library Association as "the set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating ethically in communities of learning (2016)." This goal utilizes ALA's framework to fulfill the information literacy frame, while focusing specifically on finding, evaluating and using primary sources and datasets relevant to the course.

Digital Literacy is the ability to think critically about the effect of media and web technology on communication and writing, and the ability to create scholarly content in a variety of different forms. There is a focus specifically on web publication, and methods for scholarly multimodal writing, as well as an understanding of power and influence on the web.

Data Literacy is the ability to critically use, curate, process, and produce data and data-driven analysis. This goal draws upon Data Information Literacy, a framework developed primarily for graduate students in the sciences, to teach some of the most crucial data skills, and makes them applicable for undergraduate students across disciplines (Carlson, et al. 2011).

Computational Analysis lessons are grounded in disciplinary methodology, and illustrate the ways in which scholars are using computationally-aided methods to conduct research. In addition to using computational analysis to explore and analyze sources, it also holds value in its ability to challenge students to think differently about a resource - to break down the way we convey information and think about ways to work through those abstractions.

The poster will include a description of each of the competency goals, their theoretical grounding, some examples of what specific skills or outcomes may be included in each, and an example of its application in a content-specific exercise.

In LEADR, there have been three clear benefits from early experiences with the framework. The framing is useful when speaking with faculty members who may be skeptical or unfamiliar with the Digital Humanities. Many who are skeptical of digital methods have often become familiar with Digital Humanities as a new set of methodologies and practices aimed at disrupting and displacing older methods, or as a flashy way to get students interested. Instead, by introducing course modules as a method for teaching longstanding objectives in new contexts, digital work is seen as less gimmicky, and more essential to the development of undergraduate students. The framework has also been a valuable tool for organizing teaching and instruction partnerships for the courses. The collaborative nature of digital projects and the variety of skills required calls for the involvement of scholars, researchers, and technology specialists in teaching and course management. The framework allows for more clear and open communication between partners, and clarifies the necessary skills and objectives needed. Finally, the framework has ensured that exercises are designed with student development in mind, rather than working back from a desired final project.

Bibliography

Association of College & Research Libraries [**ACRL**]. (2016). *Framework for Information Literacy for Higher Education*. http://www.ala.org/acrl/standards/ilframework (accessed 3 March 2016).

Carlson, J. R., Fosmire, M., Miller, Ch., Sapp, N., Megan, R. (2011). *Determining Data Information Literacy Needs: A Study of Students and Research Faculty.* Libraries Faculty and Staff Scholarship and Research. Paper 23. http://docs.lib.purdue.edu/lib_fsdocs/23. (Accessed 3 March 2016).

Conjuring up the Artist from the Archives: Ivar Arosenius. Digitization and Coordination of Archives for Enhanced Accessibility and Research

Mats Malm
mats.malm@lir.gu.se
University of Gothenburg, Sweden

Jonathan Westin
jonathan.westin@conservation.gu.se
University of Gothenburg, Sweden

Cecilia Lindhé
cecilia.lindhe@lir.gu.se
University of Gothenburg, Sweden

Sverker Lundin
sverker.lundin@gu.se
University of Gothenburg, Sweden

Dick Claésson
dick.claesson@litteraturbanken.se
The Swedish Literature Bank

How can our understanding of an artist be deepened and developed through digital materials and methods? How can we develop tools for a better understanding of previous practices in conjuring up, modifying and curating artists and works of art in museum exhibitions, publications and studies? What ideological and practical considerations and presuppositions have governed the presentations that have formed the artist for the public consideration?

These are questions we investigate through a three-year project about Ivar Arosenius, a Swedish artist and writer. His main body of work was produced during the last few years that led up to his untimely death in early 1909, only 30 years of age and within months of his big breakthrough. During the subsequent years and decades, his substantial production earned him recognition posthumously both nationally and internationally, and today he is one of the most renowned Swedish artists.

At the core of the project is the development of a digital archive that collects the digitized material from several sources, both public and private, into a central repository, allowing scholars and the public to view, filter, and combine the entirety in new ways, and, through public APIs that we make available, explore, activate and make use of this rich material on various platforms. In addition to this, the project has also instigated a number of studies of what knowledge and aspects can be added through different technological developments, as well as what knowledge and values are lost or threatened in a digitization process. The poster focuses on two advanced studies; the processes involved in translating a physical archive into a digital, and methods through which to give body, context, and affect back to a digitized material. In the former study, we follow the material as it travels from the manuscript vault into the digitization studio, mapping all the actants involved in shedding it of its physicality. This is a translation process that functions to rephrase the archival material with the purpose of making it mobile and conform to those protocols that define something as being digital. This rephrasing does not only remove physicality, but does also introduce a whole new vocabulary that in many ways replaces the one that art historians, archivists and conservators use to describe the manuscripts.

The latter study explores material pertaining to Arosenius' home in Älvängen, torn down in the early seventies after decades of neglect. Using the archive as a source, a virtual model is assembled in a game engine where the connection between artist, art and place is investigated to catch the way Arosenius has translated his surroundings and to contextualize the documents of the archive.

As with the digital archive, with this interactive reconstruction we aim to construct a synthesis of a heterogeneous and sometimes conflicting material that can be used both as an access-point to the life of Ivar Arosenius and his art, and as a repository: built on a source material consisting of archival photos, local stories and historic maps, paintings, 3d-scanned artefacts, sound recordings, and inventories of both the belongings of the artist and his family, and of the vegetation on his lands, the digital construction is a knowledge-model containing all the material pertaining to this part of the artists' life. As such, of central interest for the study is how to communicate interpretative practices to the user, balancing an incomplete source material with the need to create a space that can inspire affect.

Just as the archive contains a translation of Arosenius' home, first into documents and files, and later, when digitized, into bits, the reconstruction of Arosenius' home is a translation of bits into context. It is an investigation into both the limits of the **archive view**, that which the archive lets us perceive, and the produce of activating and giving depth to the archive by bringing it together with the site of its origin. In the study, we frame the act of digitally reconstructing a site as an iterative research method of investigation and translation between different media, that allow a disparate material to be collected, studied, and processed simultaneously. Arosenius' Älvängen is at the centre of this study as it is the locus of the art, and also the place of the life that the archive tries to represent. As such, it is the archaeo-archival embodiment of the artist's archive.

The project involves a number of departments and divisions at the University of Gothenburg as well as the Swedish National Museum in Stockholm, and the Museum of Art in Gothenburg.

The Latin of Matthew of Cracow (c. 1345–1410) – a corpus based study of his language and style

Jagoda Marszałek
jagoda.marszalek@ijp-pan.krakow.pl
Institute of Polish Language, Polish Academy of Sciences

The poster illustrates the problems arising when attempting to describe the language used by the author flourishing at the turn of the Middle Ages and the modern era. It shows how a corpus-based linguistic analysis can tell us more about author's language and stylethan the traditional stylometric techniques. It also presents an example of the authorship attribution research of an anonymous medieval treaty carried out with the help of digital tools.

Matthew of Cracow (c. 1345–1410) is considered as one of the greatest theologians of the Eastern Europe Middle Ages as well as one of the most important representatives of the Polish Scholasticism. He was born in Cracow, probably into a family of German ancestry. After completing his initial education in his native city, he went to Prague, where he entered the Charles University and where he became professor of theology. He was involved in the establishment of the University of Heidelberg, University of Chelmno (Poland) and later in the project of bringing to life the Academy of Cracow. He died on March 5, 1410, in Heidelberg. Matthew belonged to the followers of the *via moderna* trend of Scholasticism and is known as a zealous reformer of the Catholic Church. He is an author of more than 70 treatises, from which the most famous are *Rationale operum divinorum*, *De squaloribus Curiae Romanae* and *Opuscula Theologica* (Nuding, 2007).

The poster will present the results of the Matthew's idiolect analysis. It will focus on language change between the late Middle Ages and the early Renaissance, as one of the main goals of the study is to trace two technolects used by authors living in this period, namely Scholastic and Humanistic Latin. Scholastic Latin is defined as a variant of late Medieval Latin, the language of science and the university community in use between 12th and 14th century (Herren, 1996: 124). Its main characteristics are excessive formalism, the use of impersonal narrati on and the pursuit of semantic precision to the detriment of aesthetic and literary values. As a result, Scholastic Latin became a highly technical language, employing countless number of theological and philosophical terms (Bourgain and Hubert, 2005: 62). With the raise of the Renaissance the language of the scientific treatises underwent a significant transformation has been transformed. Imitation of the works of Virgil and Cicero made scholars to change their own way of writing. From now on they will be using the idiomatic structure, the lexis and the rhetorical and stylistic techniques characteristic – at least in their opinion – of the Classical authors (Knight and Tilg, 2015: 4; Tunberg, 1996: 130). Many of the features of Matthew's idiolect seem to follow this new tendency and it is very important to establish how big influence it had on the style of his writings.

The research is based on the corpus of Matthew's works which includes five editions published between 1930-2011 and consists of over 60000 words. It contains both his original treaties, namely *De contractibus, De squaloribus Curiae Romanae, Sermones de sanctis, Lectura super Beati Immaculati, Opuscula Theologica*, and an anonymous medieval work, often attributed to Matthew – *Ars Moriendi*.

The present study uses the methods developed by the computational stylometry and corpus linguistics and includes among others concordance, collocation and keywords analysis. The software employed for this purpose is the TXM platform – an open-source environment based on CQP and R and providing tools for NLP pre-processing, quantitative analysis together with clear metadata model (Heiden, 2010). Thanks to the TreeTagger coupled with the Latin language model file, TXM makes it possible to annotate Latin text with PoS and lemma tags, and as a result allows to perform advanced queries. In addition to frequency lists for any token property (type, lemma, PoS), the built-in CQP search engine lets generate lists for syntactic constructions specific to Classical and/or Medieval Latin, such as the accusative with infinitive or the possessive dative construction. The R package stylo (Eder et al., 2013), on the other hand, allows to trace similarities between anonymous *Ars Moriendi* treatise and the original Matthew's works.

In thatthe poster shows how the use of the Digital Humanities methods can support the traditional analysis of a writer's language and style and help to ascertain the authorship of various texts coming from different ages.

Bibliography

Bourgain, P. and Hubert, M. C. (2005). *Le latin médiéval*. Turnhout: Brepols.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts*. Lincoln (NE): University of Nebraska-Lincoln, pp. 487-89.

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *24th Pacific Asia Conference on Language, Information and Computation 2010: Conference Abstracts.*Sendai: WU, pp. 389-98.

Herren, M. W. (1996). Latin and the vernacular languages. In: Mantello F. A. C., Rigg A. G. (eds.), *Medieval Latin: An Introduction and Bibliographical Guide*. Washington DC: Catholic University of America Press, pp. 122-29.

Knight, S. and Tilg, S. (2015). *The Oxford Handbook of Neo-Latin*. Oxford University Press.

Nuding, M. (2007). *Matthäus von Krakau*. Tübingen: Mohr Siebeck.

Tunberg, T. O. (1996). Humanistic Latin. In Mantello F. A. C. and Rigg A. G. (eds.), *Medieval Latin: An Introduction and Bibliographical Guide*. Washington DC: Catholic University of America Press, pp. 128-35.

Edit Histories and Literary Turf Wars: John Ashbery, Academic Criticism and Wikipedia

Jim McGrath
james_mcgrath@brown.edu
Brown University, United States of America

How is cultural authority visualized in social media through the publication, revision, and erasure of citations? This poster highlights the social dimensions of Wikipedia's creation, revision, and dissemination. Many of us may be familiar with the mechanics of knowledge production on Wikipedia: the "citation needed" requests that follow unattributed information on Wikipedia pages, the lengthy revision histories available to interested readers, the use of bots as well as human editors, the privileging of public domain images (among other dimensions). Attention will be paid to the interfaces of the Wikipedia and its articles: their performative dimensions as sites of cultural authority, the edit histories of articles and the forms of gatekeeping revealed in patterns of page revisions, and the visualization tools created by the Wikipedia community to visualize patterns in composition and citation. More broadly, I want to situate the investments in particular patterns and performances of curation and citation present in Wikipedia within the larger context of modes of cultural production found on the web, focusing particular attention on forms of erasure and practices that, for various reasons, disavow or ignore investments in citation and attribution: screencapping, the creation of image macros, and other transformative uses of cultural objects on Twitter, Tumblr, and other platforms.

I use Wikipedia pages related to poet John Ashbery to highlight the ways ideas of literature circulate in one of the world's largest and most accessible encyclopedic resources. I am interested in Ashbery because of his literary stature in the eyes of segments of North American and global audiences. Additionally, Ashbery creates particular challenges for scholars (and other readers) interested in periodization and reception history; he has ties to several literary coteries across time and space (the New York School, L=A=N=G=U=A=G=E poetry), his work has been praised by scholars, journalists, and MTV (among other audiences), andhe continues (as of this writing) to publish new material that unsettles the work being done to cement his legacy. The various editorial and citational practices involved in creating and revising that legacy on Wikipedia can tell us much about the impact of academic and scholarly works, the privileging of certain modes of reading and cultural analysis, and the various investments in certain ideas about the value of poetry that circulate on the web. Wikipedia is an extremely visible and malleable public space where competing claims about aesthetics collide and re-collide.

Free FannyPacks: A Model for the Easy Digital Publication of Archival Periodical Material

Kevin McMullen
mcmullen.kevinm@gmail.com
University of Nebraska-Lincoln, United States of America

The nature of nineteenth-century culture, particularly literary, publication, and print culture, meant that many female writers plied their trade in periodicals. Even many American writers who we think of today primarily as novelists—Harriet Beecher Stowe, for instance—first published much of their material in periodicals. Stowe's *Uncle Tom's Cabin* first appeared serially in *The National Era*, an abolitionist newspaper. But writers like Stowe gained and have maintained notoriety in part because their work also existed in book form. Many writers whose work remained trapped in periodicals have since fallen off the literary map, their writing accessible only in increasingly fragile and scattered print runs of newspapers held in libraries and archives, or on microfilm. Attempts to digitize cultural heritage material in periodicals have been far spottier than comparable attempts to digitize books for a variety of mostly practical reasons.

Over the past decade or so, scholars such as Kenneth Price, Susan Belasco, and Meredith McGill have argued for both an increased acknowledgement of periodicals and periodical writing as a key site of intellectual and literary exchange in the nineteenth century, and the increased utilization of digital tools for the editing, study, and dissemination of periodicals.

While the advent of the digital archive has afforded well-documented possibilities for the recovery and, more importantly, dissemination of previously unknown and/or largely inaccessible material,

digital transcription, encoding, and publication of literary texts still remain skills that many academics feel are well beyond their technical capabilities. This means that many of the people best positioned to undertake such recovery work—literary scholars and other subject-specialists—are held back merely by a technological learning curve that they feel is too steep.

For the past year I have been using basic and widely-available digital tools and resources to build a digital archive of the newspaper writing of the nineteenth century American writer Fanny Fern, who, in the 1850s, was the highest-paid periodical writer in the country, writing for the widest circulated American periodical of its day. The project, *Fanny Fern in The New York Ledger* (http://fannyfern.org) is the first attempt to make available the full run of Fern's newspaper columns. Using the Drupal content management system, I have been making TEI-encoded transcriptions of Fern's columns, high-resolution digital images of the complete *Ledger* issues, and brief critical apparatuses about both Fern and the *Ledger* available for free public and scholarly use. While the digital methods and tools used to construct my project were fairly simple, the functionality and appearance of the finished product belie the relative ease (from a technical standpoint) of its creation. But the *Ledger* is just one paper and Fanny Fern is just one writer, albeit a significant one. "There are countless other periodicals and writers, particularly women writers, that deserve this sort of attention," I thought. "If only other 19th century lit scholars could see how easy this is!" And an idea was born.

I already had a relatively simple TEI template designed to handle the metadata, textual content, and linking of digital image files for periodicals. I had an XSLT style sheet that converted the TEI to HTML. I had project documentation for how I had set up my own domain name and hosting space (using Reclaim Hosting, a web hosting service specifically designed for educators and students). I had documentation about how I had installed and configured Drupal. I had documentation about how I had incorporated the HTML of the transcribed newspaper columns into the Drupal architecture. And I had documentation about how I had tweaked and played around with the design of the site. In short, I had everything anyone would need to build his or her own digital archive; all they would need is the content. If I could just share these files and instructions with other literature scholars, scholars who themselves are experts and masters of all kinds of content, then piece by piece and site by site, the gaps in literary and historical scholarship could to begin to close, be it ever so slightly. While I knew this had long been one of the (increasing number of) mantras of digital humanities, I now felt that I had the means to do my small part to contribute to its realization. Thus, using *Fanny Fern in The New York Ledger* as an example and template, I plan to soon begin providing other literature scholars with a single package, a package that contains the tools and instructions they will need to construct their own digital archives—I'm going to start handing out FannyPacks.

Geared mainly for those wishing to gather and display texts in a digital environment, these FannyPacks (a zipped collection of files) will include the needed TEI templates, style sheets, and thorough but simple documentation about digital imaging, hosting setup, and Drupal installation and execution. Scholars with a bit of tech savvy can choose to begin hosting their own projects right away. For those looking to first experiment before fully diving in, Reclaim Hosting easily allows for multiple subdomains to be hosted under an existing domain free of charge. Thus, I will provide testing space on my "fannyfern.org" domain where users can download their own installations of Drupal and experiment with adding their own content. While I have chosen to take the time to encode Fern's newspaper columns in TEI—for the preservation, interoperability, and potential for enhanced functionality of the files and their content—there are some users who might wish to simply "get their material out there" using HTML. Drupal's graphical user interface allows for the easy input of basic or full HTML. Thus, users unfamiliar with or not wanting to take the time to encode in TEI could simply encode their transcriptions in basic HTML and paste them into Drupal's GUI. The setup of both Reclaim Hosting and Drupal also provide ample opportunity for student involvement in the creation of such projects.

Projects such as *Fanny Fern in The New York Ledger* and those facilitated by the FannyPacks occupy a scholarly space somewhere between large-scale, institutionally-hosted TEI-based projects such as the *Willa Cather Archive* or *Walt Whitman Archive* (both of which are well-funded and boast a host of technical and subject specialists) and a low-cost collaborative TEI repository such as the TAPAS Project (http://tapasproject.org). While not requiring any institutional technical resources, projects based on the FannyPacks model possess greater customization and more autonomy than TAPAS projects. And, in the world of nineteenth-century American women's literature at least, there seems to be a desire for just such a model. A recently created listserv of the Society for the Study of American Women Writers (SSAWW) is geared specifically to those scholars already working on or interested in creating digital projects devoted to American women writers. And initial communication on the listserv has made clear the appeal and potential of a method and means for facilitating the publication of digital scholarly content centered around women writers, particularly periodical writers. But while my main and initial goal will be to work specifically with nineteenth century literature scholars, particularly scholars of women's literature, the FannyPacks certainly have a broader application and could be adapted to fit any manner of digital archival project.

My poster will thus provide a brief overview of my current project, *Fanny Fern in The New York Ledger*, a discussion of the project's potential to serve as a model, and specifics of the FannyPacks and their creation, distribution, and application.

Notes

[1] While there have been a number of large-scale newspaper digitization projects, most of them have been undertaken by large commercial entities that charge a fee for access to the content and provide material of mixed quality. Readex and Proquest are two of the more popular services, and genealogy

site Ancestry.com has also undertaken its own mass digitization of government records, census data, and newspaper and periodical material; all three charge for access to the content. While these services can certainly be useful for certain types of research work, the quality of the digitization, particularly transcription, is in nearly all cases quite poor, with transcriptions being derived from optical character recognition (OCR) software, which often has difficulty accurately transcribing the small and often smudged print of nineteenth-century periodicals.

[2] See Price and Belasco's introduction to their edited collection, *Periodical Literature in Nineteenth-Century America*. Charlottesville, VA: University Press of Virginia, 1995. Also, see: Belasco. " *Whitman's Poems in Periodicals*: Prospects for Periodicals Scholarship in the Digital Age." *The American Literature Scholar in the Digital Age*. Ann Arbor, MI: University of Michigan Press, 2011; McGill, Meredith. *American Literature and the Culture of Reprinting, 1834-1853*. Philadelphia: University of Pennsylvania Press, 2003.

[3] Perhaps the most relevant example of a digital recovery project, for the purposes of my presentation, is the *Women Writers Project* (http://wwp.northeastern.edu), now run out of Northeastern University and directed by Julia Flanders. Begun in 1988, the project has been instrumental in the recovery of rare or inaccessible work by early modern women writers (the project covers a period from 1526-1850, although the vast majority of texts are from the sixteenth and seventeenth-centuries). In addition to providing access to digitally encoded texts, the project has also provides various research and teaching materials. However, the *Women Writers Project* is only accessible with a paid subscription and does not deal with periodicals.

Notoriously Toxic: The Language and Cost of Hate in the Chat Systems of Online Games

Ben Miller
miller@gsu.edu
Georgia State University, United States of America

Antal van den Bosch
vandenbosch@let.ru.nl
Radboud University, The Netherlands

Cameron Kunzelman
ckunzelman1@student.gsu.edu
Georgia State University, United States of America

Jennifer Olive
jolive1@gsu.edu
Georgia State University, United States of America

Wessel Stoop
w.stoop@let.ru.nl
Radboud University, The Netherlands

Kishonna Gray
Kishonna.Gray@eku.edu
Eastern Kentucky University

Cindy Berger
cberger@student.gsu.edu
Georgia State University, United States of America

Shiraj Pokharel
spokharel3@student.gsu.edu
Georgia State University, United States of America

'Notoriously Toxic' presents a preliminary study of the language and impact of hate speech in the chat systems of online games. Developed by a group of researchers in game studies, computational linguistics, sociolinguistics, and law and guided by an overall tripartite feedback model broadly corresponding to shielding potential victims from harm, educating those who casually engage in hate speech, and censuring those who persist in abusing their fellow players, the hope is that research-driven technical and social interventions might slowly shift online discourse norms away from casual, vicious, and potentially dangerous speech. Identification at scale of textual expressions of toxic behavior in online environments is a necessary, empirical preliminary aspect of this work to understand the prevalence and cost of online hate, as is qualitative cultural studies of the games and their player populations. A recent example of qualitative framing work in this area was the 'Mapping Study on Projects Against Hate Speech Online' released in 2012 by the British Institute of Human Rights for the Council of Europe project, *Young People Combating Hate Speech in Cyberspace* (The British Institute of Human Rights Council, 2012). That report provides terminology and an environmental scan of processes aimed to limit hate speech online and offers suggestions as to new procedures. It, along with an examination of the reporting systems implemented across a host of online games, computational modeling of the language prevalent in these chat systems, and a study of work in the political sphere to defuse hate speech prior to its catalyzation of violence, serve as the foundation for this research.

Recent inquiries into the toxic elements of gaming cultures have primarily focused on communication outside of a game environment. For example, critical discourse analyses of player posts to online gaming forums found that heteronormative undertones of the World of Warcraft player community creates a culture of hostility toward LGBTQ communities (Pulos, 2013) and the same forum's adamant disavowal of feminism have made community conversations about gender roles and/or equality all but non-existent (Braithwaite, 2013). Similarly, Gray's (2012a; 2012b; 2012c) ethnography of Xbox Live demonstrates the constant barrage of gender and racially motivated harassment faced by women of color who opt to communicate with teammates via voice chat. Finally, community leaders' adamant position of gender based harassment being a 'non-issue' is summarized by Salter and Blodgett (2012), whose case study of Penny Arcade's (a popular webcomic and organizers of PAX, a successful annual gaming convention) dismissal of its responsibility in perpetuating rape culture and SXSW Interactive's recent declaration that conversations about harassment in the games space can by definition not be civil (Sinders, 2015) is indicative of an industry that is highly resistant to change unless external pressure is applied. Taken together, this scholarship is evidence that toxicity exists across gaming culture writ large, and is not isolated to a particular game or specific player community.

Studying this phenomena at webscale and in the ephemeral environments of multilingual online chat systems is complex and requires a multidisciplinary approach bridging core strengths in the humanities, such as cultural criticism, with strengths in social psychology, the data sciences, and linguistics. Studying the socially destructive behavior as manifested in online gaming platforms encourages innovative approaches to this problem. One corpus examined as part of this research is comprised of the chat logs produced by the player base in Riot Games' League of Legends (League). As of January 2014, League had ~27 million unique players every day each playing no less than 20 minutes and a peak concurrency of 7.5 million people who collectively have logged billions of hours of total play time for the game since 2009 (Sherr, 2014). Given that the game is a global phenomenon, the chat logs contain harassment in virtually every language.

Based on the UN framework provided in the International Covenant on Civil and Political Rights (1976), Susan Benesch generally defines hate speech as '. . . an expression that denigrates or stigmatizes a person or people based on their membership of a group that is usually but not always immutable, such as an ethnic or religious group. . . . Speech may express or foment hatred on the basis of any defining feature of a minority or indigenous people, such as ethnicity or religion – and can also denigrate people for another "failing", such as their gender or even their location, as in the case of migrants' (Benesch, 2014, pp. 20). This broad but inclusive definition is further elaborated upon by Nazila Ghanea in reference to the *International Convention on the Elimination of All Forms of Racial Discrimination* (ICERD) in the establishment of a spectrum from least to greatest: discriminatory speech, hate speech, incitement to hatred, incitement to terrorism, and incitement to genocide (Ghanea, 2013, pp. 940-1). The characteristics of these definitions reflect the significant impact that hate speech acts have on the establishment and enforcement of personal and communal identity and the need to identify such acts in order to preserve those identities. In his discussion on Carey's ritual model of communication as applied to cases of hate speech, Clay Calvert explains how hate speech initializes and perpetuates the subordination of one group over another (Calvert, 1997). Calvert notes that hate speech acts, specifically focusing on the repeated utilization of racial epithets, construct reality in the speaker, audience, and target members of the discourse through the creation and maintenance of mental schemata similar to the functions of other speech acts: 'In particular, racist speech helps to define who minorities are and how others think about minorities, facilitating their unequal treatment' (Calvert, 1997, pp. 12). This construction is harmful in its immediacy to the target as well as in the long-term situation as it perpetuates unequal power structures based on criteria of identity (Calvert, 1997, pp. 15-16). The construction of reality based on hate speech acts is also relevant to a discussion of online environments as the textual communication serves as a large social aspect of both on- and offline environments.

Toxicity consists of verbal expressions and behaviors that serve to destabilize groups. It is unclear whether toxic behavior is directed to elicit particular responses, and hence systemic, or reactionary, emotional, and hence, situational. Frequently, the term 'troll,' or 'trolling,' is used synonymously with toxicity. Regardless, these behaviors are necessary to address because they are a key factor in the outright hostility of online gaming environments to those perceived as other. This destabilization maps on to offline models of gender, race, class, ethnic, national, linguistic, and abelist-based hate speech. A fertile ground for this analysis is in the virtual worlds of online gaming. For example, Kou and Nardi (2013), in their research on League of Legends, found that antisocial behavior destabilizes online communities but is addressable by social code and regulatory systems.

The targeted examples in online gaming environments show a concentrated sample of toxic behavior that is pervasive in every online environment. Studies have documented antisocial behavior and toxic speech in most digital platforms (O'Sullivan & Flanagin, 2003). Whether considering flaming on 1980s and 90s USENET forums, social media fueled outrage on contemporary politics, or in-game, text-based conversation in multiplayer games, toxicity can be motivated by emotional, intellectual, political, or other causes and as such correlates strongly with the modes and consequences of offline hate speech. Understanding online toxic behavior in ways that allow for moderation of its causes and effects first requires an understanding of how to study the concrete manifestation of this behavior—the text produced by users of a system. The challenges faced by this research are many: the writing styles are heavily infused with jargon, the orthography is non-standard, the chat stream only represents one channel of communication, and the communities are fluid.

In response to these challenges, an approach grounded in machine learning and NLP was tested. Using a small subset of the available data, a classifier based upon developed to separate players toxic from non-toxic players yielded a precision of 0.77, a recall of 0.79 and an F-score of 0.78. These results are encouraging, and along with training on a larger data set, secondary factors such as player avatar gender, length of match, and others were also preliminarily tested and found to have small influencing effects. These results suggest that there are concrete, detectable semantic and syntactic patterns in the harassment levied at players in these games. Connecting these findings to mechanisms for shielding, reforming, and censuring players, and to frameworks for understanding the social and psychological costs of being effectively locked in a room with one or more individuals determined to verbally abuse a peer is the more complex task of the cultural and ethnographic studies of digital communities.

Bibliography

Benesch, S. (2014). Defining and diminishing hate speech, *State of the World's Minorities and Indigenous Peoples 2014*, pp. 14-25.

Braithwaite, A. (2013). 'Seriously, get out': Feminists on the forums and the War(craft) on women. *New Media & Society* 0(0): 1-16. First published online June 12, 2013: 10.1177/1461444813489503.

The British Institute of Human Rights Council. (2012). Mapping Study on Projects against Hate Speech Online. *Young People Combating Hate Speech Online*. http://www.coe.int/t/dg4/youth/Source/Training/Training_courses/2012_Mapping_projects_against_Hate_Speech.pdf (accessed 29 August 2014).

Calvert, C. (1997). Hate Speech and Its Harms: A Communication Theory Perspective. *Journal of Communication*, **47**(1): 4-19. First published online February 7, 2006: 10.1111/j.1460-2466.1997.tb02690.x.

Ghanea, N. (2013). Intersectionality and the Spectrum of Racist Hate Speech: Proposals to the UN Committee on the Elimination of Racial Discrimination. *Human Rights Quarterly*, **35**(4): 935-54. First published November 2013: 10.1353/hrq.2013.0053.

Gray, K. (2012a). Deviant Bodies, Stigmatized Identities, and Racist Acts: Examining the Experiences of African-American Gamers in Xbox Live. *New Review of Hypermedia and Multimedia*, **18**(4): 261-76. First published online: December 3, 2012: 10.1080/13614568.2012.746740.

Gray, K. (2012b). Intersecting Oppressions and Online Communities: Examining the experiences of women of color in Xbox Live. *Information, Communication & Society*, **15**(3): 411-28. First published online: December 19, 2011. 10.1080/1369118X.2011.642401.

Gray, K. (2012c). Collective Organizing, Individual Resistance, or Asshole Griefers? An Ethnographic Analysis of Women of Color in Xbox Live. *Ada: A Journal of Gender, New Media, and Technology* 2. First published online: June 2013. 10.7264/N3KK98PS.

Kou, Y. and Nardi, B. (2013). Regulating anti-social behavior on the Internet: The example of League of Legends. *iConference 2013 Proceedings*, Fort Worth, TX: February 12-15, 2013, pp. 616-22. 10.9776/13289.

O'Sullivan, P. B. and Flanagin, A. J. (2003). Reconceptualizing 'flaming' and other problematic messages. *New Media & Society*, **5**(1): 69–94.

Pulos, A. (2013). Confronting Heteronormativity in Online Games: A Critical Discourse Analysis of LGBTQ Sexuality in World of Warcraft. *Games and Culture*, **8**(2): 77-97.

Salter, A. and Blodgett, B. (2012). Hypermasculinity & Dickwolves: The Contentious Role of Women in the New Gaming Public. *Journal of Broadcasting & Electronic Media*, **56**(3): 401-16.

Sherr, I. (2014). Player Tally for *League of Legends* surges. *The Wall Street Journal*. First published January 27, 2014. http://blogs.wsj.com/digits/2014/01/27/player-tally-for-league-of-legends-surges/.

Sinders, C. (2015). I Was On One Of Those Canceled SXSW Panels. Here Is What Went Down. *Slate*. First published October 29, 2015. http://www.slate.com/articles/double_x/doublex/2015/10/sxsw_canceled_panels_here_is_what_happened.html.

United Nations General Assembly. International Covenant on the Elimination of All Forms of Racial Discrimination. *United Nations*. http://www.ohchr.org/EN/ProfessionalInterest/Pages/CERD.aspx (accessed 4 August 2014).

El Uso de las Tecnologías Digitales para la Difusión del Patrimonio Cultural en México

Ernesto Miranda Trigueros
mirandatrigueros@gmail.com
Instituto Nacional de Antropología e Historia (INAH)

Ernesto Priani Saisó
epriani@gmail.com
Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México - UNAM, Mexico

Isabel Galina Russell
igalina@unam.mx
Instituto de Investigaciones Bibliográficas, Universidad Nacional Autónoma de México - UNAM, Mexico

Esta presentación describe el proceso de creación de la Red de Tecnologías Digitales para la Difusión del Patrimonio Cultural (RedTDPC), cuyo objetivo es generar conocimiento interdisciplinario y buenas prácticas en la aplicación de tecnologías digitales para la difusión del patrimonio cultural en México.

Antecedentes

La digitalización del patrimonio cultural es fundamental para su difusión a nivel mundial[1]. Para los países en desarrollo que tienen un importante patrimonio cultural, existen numerosos obstáculos en términos de infraestructura y gestión. En México, uno de estos obstáculos es la conceptualización de lo digital como un ámbito secundario dentro del sector cultural. Aunque existe una creciente mención de la importancia de trasladar los contenidos culturales a medios digitales, es muy poco lo que realmente se traduce a la práctica (Priani, 2012). Por otra parte, han existido proyectos aislados, pioneros en la digitalización del patrimonio[2] y organizaciones como la RedHD, que han servido como nodos para discutir estos temas. En el 2014 se llevó a cabo el 1er Congreso de Patrimonio Cultural y Nuevas Tecnologías organizado por el INAH donde se mostró que si bien México desarrolla aplicaciones de tecnología para la preservación e investigación del patrimonio hay poco trabajo en temas de difusión.

Creación de la RedTDPC

En este marco y atendiendo a una convocatoria del CONACYT,[3] se creó la RedTDPC en abril de 2015, con un programa académico encaminado a identificar actores[4] y encontrar sus temas en común. Los temas que se han identificado hasta el momento son:

- La importancia de debatir políticas públicas, tipos de licencias y derechos de autor.
- Las relaciones no siempre equitativas entre las industrias creativas y las instituciones
- La pertinencia de hacer investigación de campo previa al desarrollo de proyectos, dada la diversidad de México en términos de accesibilidad.
- Políticas y buenas prácticas de digitalización, preservación y difusión de objetos culturales.

Discusión

Si bien la creación de la Red se entiende como un logro en términos de gestión institucional, es necesario resaltar las dificultades para su desarrollo:

1. Falta de coherencia entre requisitos institucionales y objetivos de la Red

El CONACYT hace hincapié en que las redes deben apoyar el 'desarrollo económico y tecnológico'[5], sin embargo articular la importancia de la difusión del patrimonio cultural en estos términos es complicado. En muchos sentidos, la creación de esta Red ha permitido pensar en un espacio que equilibre en términos humanísticos esa exigencia. Al mismo tiempo, la Red se formó dentro de un programa que establece procesos burocráticos a los que es necesario apegarse, pero que no están definidos institucionalmente, lo que ha conducido a que la toma de decisiones sea lenta y se dedique mucho tiempo a temas operativos de la Red, restándole tiempo a la discusión de los temas sustantivos, algo que eventualmente podrá cambiar.

2. Ausencia de definiciones conceptuales

La propia definición de "difusión" es problemática y ha merecido parte de la reflexión de la Red. La RAE[6] define esta palabra como: 'propagar o divulgar conocimientos, noticias, actitudes'. Es la palabra divulgar, la que a veces se encuentra en oposición a difusión: la primera sólo dedicada a los pares académicos, y la divulgación al público en general (Gándara, S/F).[7]

Otra problemática es la poca documentación que existe en México sobre cómo trabajar en la difusión del patrimonio con herramientas digitales. Son pocas las instituciones mexicanas que cuentan con áreas dedicadas a la experimentación y acumulación de conocimiento en torno a estas prácticas. Casi siempre, son empresas privadas las que realizan los proyectos. Esto no es negativo en sí porque permite que las industrias creativas encuentren una fuente de trabajo. Lo que sí es necesario es que estas empresas regresen esa reflexión y experiencia a las instituciones.

3. Prospectivas

Hasta ahora se ha trabajado en la organización de encuentros académicos, que son el tipo de actividades que pueden llevarse a cabo fácilmente dentro del modelo institucional en el que se creó la Red. La mayor dificultad para el futuro es fortalecer la reflexión en temas como: legislación, accesibilidad, públicos, documentación, sustentabilidad, crowdsourcing, y cuyo trabajo se traduzca en acciones concretas en esos ámbitos. Es decir, pasar de responder a las necesidades institucionales para iniciar un proceso de producción de conocimiento.

Trabajo a futuro

El siguiente objetivo de la Red es lograr que estas reflexiones se traduzcan en proyectos específicos de difusión a través de nuevas tecnologías para poder ampliar el conocimiento a través de la experimentación.

Dirección de la RedTDPC: http://redtdpc.inah.gob.mx/

Bibliography

Gándara, M. (in press). De La Interpretación Temática a La Divulgación Significativa. In Gándara, M. and Jiménez, M.A. (eds), *Interpretación Del Patrimonio Cultural*, ENCryM/INAH.

Priani, E. (2012). Finding support for disruption:developing a digital humanities project in Mexico. *Aslib Proceedings: New Information Perspectives*, **64**(1): 97-103.

Notes

[1] Ver por ejemplo, Europeana (http://pro.europeana.eu/about-us/) o CHARISMA (Cultural Heritage Advanced Research Infrastructures http://www.charismaproject.eu/) o DPLA (Digital Public Library of America http://dp.la/) o la API del Rijkmuseum (https://www.rijksmuseum.nl/en/api).

[2] Ejemplo de ellos son, la Biblioteca Digital del Pensamiento Novohispano (www.bdpn.unam.mx), Gran Diccionario Nahuatl (http://www.gdn.unam.mx/), Hemeroteca Nacional Digital (www.hdnm.unam.mx), Primeros Libros (http://primeroslibros.org/), Repositorios de Archivos y Fondo Antiguo (http://www.remeri.org.mx/archivos/), o el Archivo Digital Flores Magón (http://archivomagon.net/), por mencionar algunos.

[3] Consejo Nacional de Ciencia y Tecnología es el organismo nacional encargado de impulsar y fortalecer la investigación científica. Las redes temáticas tienen el objetivo de ampliar las redes de investigación entre talentos mexicanos como agentes de innovación, desarrollo económico y tecnológico en México

[4] Actualmente la red cuenta con 65 miembros activos de universidades, instituciones públicas y privadas, ONG's y sociedad civil.

[5] Bases de las Redes Temáticas del CONACYT.

[6] Real Academia Española

[7] En el ámbito de la ciencia en México hay una amplia discusión sobre la diferencia entre divulgar y difundir que, sin embargo, responde a objetivos distintos a los del patrimonio cultural

Challenges in Setting up a Digital Humanities Centre in Romania

Corina-Elena Moldovan
corimoldovan@gmail.com
Babes-Bolyai University, Romania

The Transylvania Digital Humanities Centre (DigiHUBB) is an emerging Digital Humanities Centre based in the Babes-Bolyai University of Cluj-Napoca, Romania. The Centre was officially established in April 2014 within the Faculty of Letters. As is evidenced in other cases, DigiHUBB exists because of the commitment of scholars experienced in various fields related to Digital Humanities.[1] The primary goal of DigiHUBB was to create a collaborative network connecting DH activities within the University and beyond, that first took into account the importance of communicating the vast global impact that DH research and practices have today. The centre defined itself as impact-oriented, with an emphasis on innovation related to research, teaching and creative activities. While following the path of other prestigious DH Centres around the world, DigiHUBB aims to develop its own expertise coming from the distinctive regional specificity of Transylvania that benefits from a multicultural and multilingual cultural heritage, which has not been sufficiently explored with modern tools till now.[2]

The importance of analysing the different ways in which a DH centre is built was stressed by Claire Warwick in her seminal essay "Institutional models for digital humanities" published in the *Digital Humanties in Practice* (Warwick, C., Terras, M., Nyhan, J., 2014). The cases analysed in the chapter - the creating of the Center for Digital Research in the Humanities at the University of Nebraska-Lincoln and the Digital Humanities in Mexico - were considered representative for the way they illustrated the institutional environment and the particularities of diverse "particular academic, political, cultural and economic realities" (http://blogs.ucl.ac.uk/dh-in-practice/chapter-9/).

This paper is a commentary on the manner in which a DH Centre is established in an Eastern-European country[3] and the challenges that exist within the traditional structures and mentalities at different levels -- the academia, the policy-makers and the society in the larger sense. The poster will focus on the strategies that DigiHUBB has, at different levels – education, research and innovation, collaboration with the IT industry. It will also point out similarities between other post-revolutionary experiences and learn if there could be a kind of specificity that could give our centre a uniqueness that will make it relevant internationally. We will compare this challenge with other interdisciplinary endeavours in Romania. The argument in this paper strongly resonates with the thematic implications addressed in this conference – the "social, institutional and multicultural aspects of digital humanities" (http://dh2016.adho.org/cfp/).

In approaching this report we have observed several paradoxes that could easily exemplify the challenges that an innovative approach faced in both the institutional environment and the investment in the research framework in Romania. In the first place, the rapidity with which DigiHUBB gained international support and assistance[4] was tempered by the resistance of the existing research structure, although it is well-known that Romanian research in general is less internationalized and under performing in all major university rankings.[5] In the second instance, although the IT industry in Romania, and especially in Transylvania, is productive and lucrative, it addresses the quotidian aspects of digital work and not the innovative part. Moreover, there is a visible mismatch between the skills needed by the knowledge market and the qualifications provided by the academia.

The challenges that DigiHUBB faces are complex and they target, as our paper will show, several important issues that go from the generic ones (the misunderstanding of the concept, the lack of confidence in its epistemological value, the supremacy of the published paper book over electronic publications, the absence of systemised pedagogy in Digital Humanities, the minimal funding from the

University or governmental institutions) to more specific ones, which are more challenging to engage with.

DigiHUBB has been involved in a variety of activities related to the Digital Humanities since its inception: in January 2014 the keynote speaker for the inaugural conference was professor Susan Schreibman from Maynooth University; this was followed by a conference held by doctor Julyanne Nyhan, from UCL, in March and many informal meetings with other key members of the Digital Humanities community. The lobbying and promoting Digital Humanities also included participation to national academic events, publishing articles on the theme[6]; the following step was to set up a training event, so in April-May 2015 DigiHUBB organized a one week workshop on TEI and data visualisation financed by NeDIMAH and the European Science Foundation, and co-organized a symposium on textual digital analysis within the Babes-Bolyai University.[7] There was a strong involvement in networking and studying the activities of other DH centres in Europe. We also succeeded to introduce an MA module on scholarly edition that will start in 2016.

In the process of setting up our centre we had to fight a very specific way of relatedness that still characterizes post-revolutionary Romanian society, which we can be characterized as "non-digital humanist", as it is, in most cases, individualistic, speculative, traditional, and change-resisting, in comparison with the collaborative, hands-on, innovative and co-creative features of Digital Humanities[8].

Understood as a HUB, our centre has multiple possibilities of offering innovation, at all levels of the research-economic-social chain. We also consider it as a new cultural model, a real co-working space opened to all the actors of our multicultural and multilingual region.

Bibliography

Byrne, T. and Schreibman, S. (2015). *'Digital Humanities and the Innovation Ecosystem: A DARIAH-Ireland Report.'* Maynooth University.

Schreibman, S., Siemens R. and Unsworth J. (eds). (2014). *A Companion to Digital Humanities*. Hoboken: Blackwell Publishing.

Terras, M., Nyhan, J. and Vanhoutte, E. (eds). (2013). *Defining Digital Humanities: a Reader*. UK: Ashgate.

Ridge, M. (2014). *Crowdsourcing our Cultural Heritage*. UK: The Open University.

Warwick, C., Terras, M. and Nyhan, J. (eds). (2012). *Digital Humanities in Practice*. Facet.

Notes

[1] Some of the DigiHUBB's members were directly involved in Digital Humanities, like text editing or mapping, others expressed their interest in Digital Humanities as a discipline and an object of study. The DigiHUBB members come from different fields, like linguistics, geography, literature, computer science, art history, etc.http://centre.ubbcluj.ro/digihubb/.

[2] There is a lack of collaborative projects concerning big historical events important for Transylvania and Romania, as the "1918 Centenary" where DigiHUBB could evidently have an important contribution.

[3] Romania is, at the moment of this proposal, one of the last East-European country that entered the Digital Humanities circuit.

[4] Immediately after its creation DigiHUBB was put on the "Around DH" map (courtesy to Alex Gill from Columbia University Libraries), http://www.arounddh.org/. The response of the Digital Humanities Community to the first call for assistance launched on Humanist Discussion Group was impressive.

[5] As reported in http://ec.europa.eu/research/horizon2020/pdf/country-profiles/ro_country_profile_and_featured_projects.pdf#view=fit&pagemode=none.

[6] See Corina Moldovan, *A discipline of reference in present days research digital humanities* , in "Globalization, Intercultural Dialogue and National Identity", Iulian Boldea ed., Arhipeleag Press XXI, 2014, pp.286-294. Corina Moldovan, *Les humanités numériques, une provocation*, in "Debates on Globalization. Approaching National Identity through Intercultural Dialogue", Iulian Boldea ed., Arhipeleag Press XXI, 2015, pp.99-108.

[7] „Editing texts in a digital world:text encoding and data visualization", workshop, Cluj-Napoca, 27 april-3 may 2015 and „Explorations in textual digital analysis for the humanities and social-sciences", symposium, Cluj-Napoca, 14-15 may 2015.

[8] „The digital humanities, therefore, not only widens the scope and processes of disciplines within the university, but contributes to national innovation agendas, creating new possibilities for the traditional scholar within an increasingly competitive academic and economic context. As such, the collaborative nature of digital humanities research contributes to the innovation ecosystem, understood as the productive interaction between people, ideas, flows, processes and outputs." (Byrne, T. and Schreibman, S. (2015).

Collecting Judgments on Artworks Through a Similarity Game

Giovanni Moretti
moretti@fbk.eu
Fondazione Bruno Kessler, Italy

Sara Tonelli
satonelli@fbk.eu
Fondazione Bruno Kessler, Italy

Rachele Sprugnoli
sprugnoli@fbk.eu
Fondazione Bruno Kessler, Italy

We present PAGANS (Playful Art: a GAme oN Similarity)[1] a playful activity to be performed by pairs of users in order to collect similarity judgments about artworks. The final goal of this task is to have indicators concerning how people perceive artworks and how they judge their similarity. We are also interested in comparing such judgments with the opinion of art curators, and see whether users' contribution can be integrated in the arrangement of a virtual or physical exhibition in view of a crowd-curation approach (Ridge, 2014).

PAGANS foresees the involvement of a pair of users at a time, who play in parallel. Each of them plays the same game independently: similarity was not explained and participants were asked to follow their intuition. A final score, presented as a sort of "aesthetic affinity score", is obtained by comparing the two judgments and how much they overlap. The game could be played online, but so far the collection method has been tested in real-world scenarios, where both players are physically in the same place and one researcher is available to give feedback after the completion of the activity.

The game itself is as follows: a virtual card representing an artwork is given (the card with a red pin on the right of the table in Fig. 1), while a set of other 10 cards is displayed to the user. (S)he has to drag and drop on the round target the cards in order of similarity to the given card, until all images on the table are ranked.

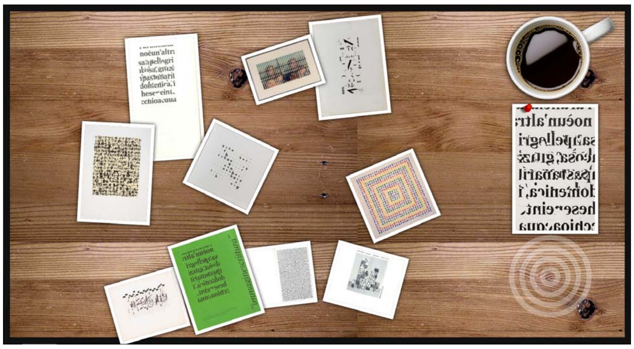
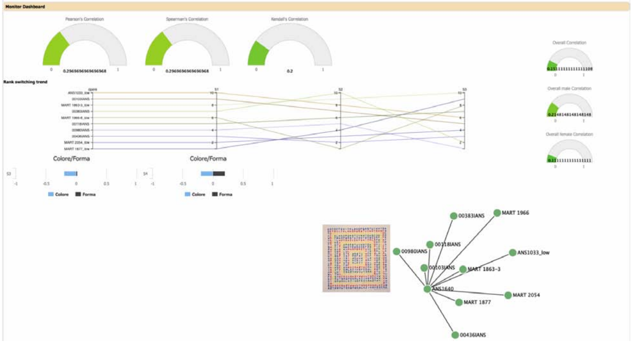


Fig. 1: Game interface

When both players have completed the task, they enter information about gender and age, and then the system shows the dashboard displayed in Fig. 2. The Pearson's, Spearman's and Kendall's coefficients (Hauke and Kossowski, 2011) are three metrics that measure with slight differences the players' agreement on the similarity judgments (the higher the value, the higher the agreement). This score is presented to the players as their aesthetic affinity score. The best affinity is reached when the players choose the same ranking, since there is no 'gold standard' order. Players' affinity is also compared with the ones displayed on the right of the dashboard: the average score obtained by other pairs previously participating in the game ("Overall correlation"), the average agreement among all male players, and that of female players. Another useful information is the "Rank switching trend": for each artwork to be ranked, the picture shows if the two players put them in the same order (straight line) or if they switched some positions.

Each picture displayed in the game was pre-processed with the LIRE tool[2] that extracts automatically image-related features such as color and shape. These features are used to provide information about the similarity judgments provided by the players, specifically if similarity relies more on color or shape information. Finally, the system outputs on the fly a network, where each node is one of the artworks in the game and the distance from the central node (i.e. the pinned card) is proportional to the average rank assigned by the two players.



Monitor Dashboard

Fig. 2: Players dashboard

PAGANS was presented during Researchers' Night 2015 in Trento (Italy) as a game for pairs of players, allowing us to collect around 170 game sessions in few hours. The game environment proved successful in engaging players also thanks to some gamification strategies. For instance, every hour the system automatically displayed a message assigning two free museum entrances to the players currently involved in the game. Besides, we kept track of the best affinity scores, and we identified the "winning pair" of the night. This boosted competition, with players trying to beat the highest score.

The goal reached with this first experiment was twofold: on the one hand, some verbo-visual works from

"Archivio di Nuova Scrittura" (Ferrari, 2012) were displayed in digital form for the first time, reaching an audience that would not necessarily see them in an exhibition. Since the game included four possible similarity sessions, around 50 artworks were shown. These works are usually kept in the archive of MART[3] and MUSEION[4] museums, not visible to the public. A second advantage of PAGANS is that we were able to collect in a short time several similarity judgments, which will be used to investigate which features related to images and possibly to persons' age and gender correlate best with similarity. These analyses are currently in progress.

Bibliography

Ferrari, D. (2012). *Archivio di Nuova Scrittura Paolo della Grazia. Storia di una Collezione/Geschichte einer Sammlung*. Silvana Editoriale.

Hauke, J. and Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, *30*(2): 87-93.

Ridge, M. M. (Ed.). (2014). *Crowdsourcing Our Cultural Heritage*. Ashgate Publishing, Ltd.

Notes

[1] https://youtu.be/PgiZl6noPns?t=5m29s

[2] http://www.lire-project.net/

[3] http://www.mart.trento.it

[4] http://www.museion.it

A Workflow for Encoding and Publishing Inscriptions

Elli Mylonas
elli_mylonas@brown.edu
Brown University, United States of America

Scott DiGiulio
scott_digiulio@brown.edu
Brown University, United States of America

US Epigraphy (USEP)(Bodel, 2015), directed by John Bodel at Brown University, is a venerable project – its roots are in a printed handbook, which was instantiated as an HTML based website at Rutgers in 1997 (Bodel et al., 1997). In 2003, Prof. Bodel moved to Brown and USEP was converted to a more automated, data-driven site. It has gone through several implementations since that time, following the state of the art in humanities computing/DH and conforming to the practices of our library development team.

The project's goal is to identify, document and create new editions of classical inscriptions in American collections, as well as to teach digital epigraphy. The inscriptions are organized geographically by collection. The project collaborates with international epigraphic consortia such as Eagle (EAGLE, 2015) and will remain conformant with their encoding schemas and vocabularies. In addition, it is one of the active contributors to the Epidoc schema for epigraphy (Elliott et al., 2015).

The current USEP front end is written in Django, and is powered by a SOLR index. XML transformations occur at two points: when the inscriptions are ingested into SOLR, and when they are displayed - the latter transformation takes advantage of SaxonCE and therefore takes place in the user's browser.

Our choice of Django for a front-end framework was driven by the environments used by our library developers – it means that we need a developer to work on the project, either library staff or a skilled student programmer. Our choice of Saxon CE to generate our display was driven by our need to modify the display easily outside of the development environment and is easily replaced.

Inscriptions are added to the collection as they are identified, with minimal information (location and a bibliographic reference or "unpublished") and iteratively enriched. Although major collections are represented, new inscriptions are always being found in small collections or in storerooms. Gradually, all inscriptions in USEP are edited to have detailed metadata, a full transcription, more complete bibliographic information and images.

The project has received some funding over the years, but has functioned continuously with basic university support. Staff include

- a faculty director
- technical management and consulting, as well as programming support provided by the library
- a graduate student/postdoc manager in charge of the encoding workflow and proofreading (along with the director) and other improvements
- graduate student encoders and collaborators outside Brown, who either contribute editions or whose students contribute editions.

Almost all USEP encoding is done by newly minted encoders or colleagues who aren't working on it consistently, it is important to make the process easy and as error-free as possible. The project is also used in digital epigraphy seminars, and the Oxygen forms allow graduate students to focus on their epigraphic work while also engaging with the decisions and activities of text encoding. We also want to make it easy for the manager to add inscriptions to the website and to modify the display and controlled vocabularies.

The components we have developed to satisfy these criteria are

- Forms in Oxygen Author mode with controlled vocabularies and proofreading transformations to assist encoders. USEP uses the Epidoc schema and a modified version of its XSL/CSS files.

- An Oxygen framework configured as an add-on to disseminate the forms, to propagate updates automatically, and to be accessible beyond Brown.
- Controlled vocabularies and shared bibliography are stored on the web, so encoders are always using the latest version.
- GitHub to store our working corpus, images, XSLT stylesheets, bibliography and controlled vocabularies. Currently, GitHub functions as our public repository, but we intend to store all USEP sources in the Brown Digital Repository when the editions are more stable. The USEP data repository on github also hosts XSL for our display transformation.
- Server side Git scripts to automatically initiate the SOLR process when any changes are committed.

Currently, encoders are trained by the encoding manager and technical manager. Their work is proofread for epigraphic accuracy by the project director and encoding manager and screened for XML validity by the encoding manager and the technical manager. Once initial encoding is done, any further corrections or editing are handled by project members. The encoding manager pushes the finalized inscriptions to GitHub, triggering the SOLR indexing process, which then makes the new inscriptions live on the Usepigraphy website. These last few steps are an iterative process, as it is important to publish inscriptions quickly, and our process makes it easy to correct mistakes. Because the XSLT and CSS are also treated as data, the encoding manager can modify the inscription display as well which makes changes to the display much more nimble than if the project had to rely on the developers.

Bibliography

Bodel, J. and Tracy, S. (1997). *Greek and Latin Inscriptions in the USA. A Checklist.* NY.

Bodel, J. (ed). (2015). *US Epigraphy Project Website*. http://usepigraphy.brown.edu(accessed 6 March 2016).

Elliott, T, Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S., et al. (2007-2015). *EpiDoc Guidelines: Ancient documents in TEI XML (Version 8)*. http://www.stoa.org/epidoc/gl/latest/. (accessed 6 March 2016).

Europeana Network of Ancient Greek and Latin Epigrapy (EAGLE). http://eagle-network.eu(accessed 6 March 2016).

UpCASE – A Web Application for Building and Maintaining Language Resources

Claes Neuefeind
c.neuefeind@uni-koeln.de
University of Cologne, Germany

Francisco Mondaca
f.mondaca@uni-koeln.de
University of Cologne, Germany

Mihail Atanassov
matanass@uni-koeln.de
University of Cologne, Germany

Introduction

UpCASE (Upload, Correct, Annotate, Search and Export) is an open source web application[1] that provides researchers and lay people a wide range of options for viewing, evaluating, editing and enriching text collections. It is conceived as a multifunction web application where users can upload text documents in different formats, including scanned texts whose characters are automatically recognized by Optical Character Recognition (OCR). While being able to search the uploaded text, users can correct, annotate and also export it into different formats.

Motivation and Background

UpCASE is the result of years of work with Romansh texts and lexical resources. Romansh[2] is the smallest of the four national languages in Switzerland with approximately 50.000 native speakers (Furer, 2005). Despite several actions by official organizations, e.g. the ratification of the *European Charter for Regional and Minority Languages*,[3] Romansh is on a continuous retreat and is by now considered an endangered language. The main motivation of our work was to create a suite of tools to support the Romansh language community in building and maintaining language resources for Romansh. However, UpCASE is not restricted to be used within this particular context.[4] Our interest here is to show the relevance of using tools like UpCASE for small languages in general, since in the European Union alone there are more than 100 languages, many of them having little or no support by official institutions.[5]

We present UpCASE together with a specific historical text collection, namely the *Romansh Chrestomathy* (RC) compiled by Caspar Decurtins (Decurtins, 1888-1919). The RC comprises texts from four centuries reflecting the different idioms of Romansh[6] and is recognized as the most important historical text collection of the Romansh language (cf. Egloff and Mathieu, 1986: 7). It contains approximately 7500 pages covering a wide range of different topics, text types and genres, and therefore is an excellent basis for the compilation of a text corpus. All in all, the RC can be seen as a monument for language, speakers, and culture of Romansh in Switzerland, and as such constitutes an exception for small linguistic and cultural communities (Rolshoven, 2012).

The RC text corpus was created in two successive projects funded by the German Research Fund (DFG). In the first project, the RC was digitized and its characters recognized by OCR. The main objective was to correct the OCR output to provide a digital full text version of the RC. Due to the characteristics of the RC as a multilingual historical text collection of a small language, with varying orthographical standards and almost no digital lexical resources available, the correction of OCR errors could not be solved in a fully automatic manner (Rolshoven, 2012). Instead, we implemented a web-based editing tool allowing native speakers to participate in the task of OCR correction.[7] In a follow-up project the corrected texts were enriched with part-of-speech (POS) tags. As in the treatment of OCR errors, a fully automatic POS-tagging procedure could not be applied to the RC (Neuefeind, 2013). First we compiled a lexical resource by digitizing lexica and generating inflected word forms. On this basis, we approached the linguistic annotation with a semi-automatic procedure, combining lexical lookup (resulting in mostly ambiguous tags) with manual correction and supervision, thus adapting the collaborative methodology from the DRC-project (Mondaca and Atanassov, 2016).

Features and Technical Aspects

UpCASE brings together the experiences of both projects, combining the key features of collaborative corpus construction, enrichment and maintenance in a single web application. While existing tools mostly focus on a particular use case like collaborative correction (e.g. Wikisource[8]) or collaborative annotation (e.g. WebAnno[9]), we connect these functionalities in a full workflow from raw to enriched text. UpCASE is based on several state-of-the-art web technologies such as Spring WebMVC, Spring Data, JAXB, and JQuery. The data is persisted in a document-oriented NoSQL database (MongoDB), whose records are structurally similar to JSON objects. The use of JSON enables a straightforward communication with other resources. Using predefined REST (Representational State Transfer) interfaces, distributed language systems may be used for data enrichment, without the need of complex data conversion.[10]

Using robust and scalable software on server side, and lightweight, clean and interactive components on client side, UpCASE offers different views in order to improve its usability. There are options to treat the collection as a whole, e.g. for searching, statistics and exporting, or to modify the data at hand, e.g. to edit, annotate or enrich. After importing text documents (or scanned images of texts which are OCR'ed), the text is indexed with Lucene and made accessible through an editable directory tree together with a full text search access. The stats view offers some basic statistical information about the text collection. In the export view, the user can choose different formats, e.g. plaintext or XML, to export the whole collection or parts of it. At the document level, each token is represented by a clickable widget, which opens a modal window containing different views – depending on user rights – associated with specific functions, e.g. editing, correction or annotation. The edit view allows the user to modify the text, e.g. to correct errors produced in the OCR process. The view contains both the editable word form and the relevant part of the scanned image with its highlighted position. The annotation view allows the user to create annotations like POS-tags on the fly, thus allowing complex searches on the search view.

Summary

Our presentation gives an overview of UpCASE and its basic functions, focusing on features for corpus maintenance, extension and enrichment. While in the first place we present a Romansh language resource, the concepts and features of this use case can be transposed to other text collections and languages. Thus UpCASE can be seen as an approach to technologically and methodologically support the preservation of the cultural heritage of regional and minority languages.

Bibliography

Decurtins, C. (1888-1919). *Rätoromanische Chrestomathie*, **13**, Erlangen: Junge. Reissued by Octopus-Verlag/Società Retorumantscha, Chur (1982-1986).

Egloff, P. and Mathieu, J. (1986). *Rätoromanische Chrestomathie: Register*. Chur: Octopus-Verlag/Società Retorumantscha.

Furer, J.-J. (2005). Eidgenössische Volkszählung 2000: Die aktuelle Lage des Romanischen. Neuchatel: Bundesamt für Statistik. http://www.bfs.admin.ch/bfs/portal/de/index/news/publikationen.Document.66870.pdf (accessed 21 February 2016).

Gross, M. (2004). *Romansh: Facts & Figures*. Chur: Lia Rumantscha.

Holley, R. (2009). Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers. National Library of Australia. http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf (accessed 21 February 2016).

Liver, R. (2010). *Rätoromanisch: Eine Einführung in das Bündnerromanische*. Tübingen: Narr.

Mondaca, F. and Atanassov, M. (2016). ARC: Annotierte Rätoromanische Chrestomathie. *Atti del VI Colloquium retoromanistich Cormons 2014*. Udine: Società Filologica Friulana, pp. 13-28.

Neuefeind, C. (2013). The Digital Romansh Chrestomathy. Towards an Annotated Corpus of Romansh. In: Zampieri, M. and Diwersy, S. (Eds.), *Special Volume on Non-Standard Data Sources in Corpus-Based Research* (ZSM Studien 5). Aachen: Shaker pp. 41-58.

Rolshoven, J. (2012). Die Digitale Rätoromanische Chrestomathie. *Ladinia*, **36**: 119-51.

Notes

[1] Code and license can be found at https://github.com/spinfo/upcase. UpCASE is delivered as Maven-based Java-project, with an embedded servlet container and database. To install and run the tool, only Java and Maven are needed.

[2] The official denomination for Romansh according to the federal constitution of Switzerland is Rhaeto-Romanic. In this paper, the term Romansh denotes Rhaeto-Romanic as spoken in the Canton of Grisons (Liver, 2010).

[3] http://www.coe.int/t/dg4/education/minlang/

[4] For UpCASE is fully open source and can be used as API, it can be extended to be used with languages other than Romansh, e.g. by implementing extensions for language-specific issues (like different charsets, encodings, lexical resources, etc.).

[5] http://www.eurominority.eu

[6] Romansh is not just one single and consistent language. In fact, it can be subdivided into 5 main idioms, namely Sursilvan, Sutsilvan, Surmiran, Puter and Vallader, all of which have an independent orthography (Gross, 2004: 13). Alongside, in 2001 the Canton of Grisons introduced Rumantsch Grischun as official literary language meant to canopy the individual idioms.

[7] http://www.crestomazia.ch. For comparable approaches at a larger scale see (Holley, 2009), among others.

[8] http://www.wikisource.org

[9] https://webanno.github.io

[10] To demonstrate the possibility to interact with a remote resource, we added a customized translation feature for the particular use case presented here. We make use of the Pledari Grond (PG), an online dictionary for Romansh that we have developed in collaboration with the Lia Rumantscha (http://www.liarumantscha.ch). A web service delivers a list of translations for a selected token by directly querying the PG. When no results are returned, the user can optionally enable a notification supplying the PG editorial staff with a request for translational support.

Medialatinitas.eu. Towards Meaningful Integration and Retrieval of Resources for Medieval Latin

Krzysztof Nowak
krzysztofn@ijp-pan.krakow.pl
Institute of Polish Language, Polish Academy of Sciences

Bruno Bon
bruno.bon@irht.cnrs.fr
Institut de recherche et d'histoire des textes, CNRS, Paris, France

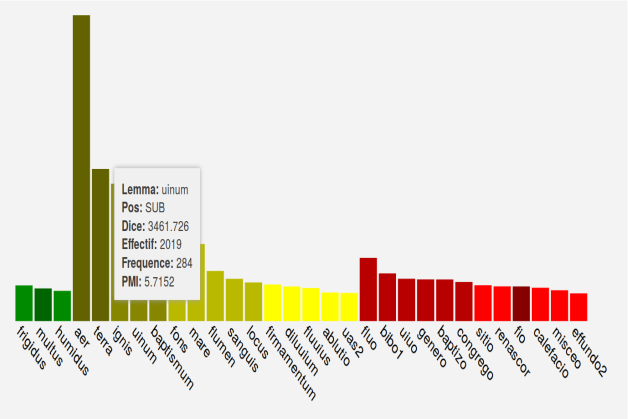
Renaud Alexandre
renaud.alexandre@irht.cnrs.fr
Institut de recherche et d'histoire des textes, CNRS, Paris, France

Latin was one of the most widely used languages in European history. In its spoken and written it was the language of daily communication, law, literature, and science for over fifteen centuries on the territory stretched from Spain to Germany to Poland and from Sweden to Croatia to Italy. The geographical, chronological and functional variation is reflected in a large number of texts which, in turn, gave rise to a vast body of secondary literature. These multifarious resources, though, even if by now partly available in digital form, remain still widely dispersed and do not easily lend themselves to integrated search.

medialatinitas.eu (Nowak and Bon, 2015) is a web mashup which aims at meaningful integration of textual, lexicographic and encyclopaedic resources for Latin. Apart from improving access to the data, the main goal of the presented application is to challenge the separation of linguistic competence and real-world knowledge in vocabulary description, as both components should effectively cooperate in comprehension of the Medieval Latin text and culture. The *medialatinitas.eu* may also compensate for major deficiencies of the resources (separate electronic text collections, for instance, covering only small proportion of the texts preserved etc.), as well as of the poorly designed interfaces and query engines they are made available through.

The heterogenous content (both academic and community-based dictionaries, thesauri, gazetteers; corpora and text collections; encyclopedias, document and image repositories, library catalogues etc.) has been interlinked only to the degree needed for its unified query. As a result, the data integration takes place mainly at the level of the web interface which thus constitutes presentational layer and a point of access to the services running in the background. When first visiting the page, users (be it lexicographers, linguists, historians etc.) come across a basic autocomplete search field: here, they can ulate the query phrase (as for now only lemma search has been implemented) which is next processed and despatched to both locally and distantly running services. The results are subsequently returned and displayed on the main page as a set of separate widgets, each of which may contain a concordance, a table, or another of data visualisation (timelines, charts, maps, lists etc.). As a whole, the widgets contribute to extensive description of linguistic and cultural properties of the lexical units.

The *medialatinitas.eu* attempts to address the drawbacks of popular dictionary aggregators in which the very fact of juxtaposing multiple resources seems often to suffice as their raison d'être. Destined for scholarly users, the *medialatinitas.eu* will make a heavy use of graphical hints and narrative devices such as interpretative notes and explicative commentaries which will accompany visual data representations in particular. On click, every widget will provide an interested user with fuller description of selected semantic or distributional properties of the word and constitute an entry point to an instance of CQPWeb (corpora), eXist-db (dictionaries), or R dashboard.


Lemma: uinum
Pos: SUB
Dice: 3461.726
Effectif: 2019
Frequence: 284
PMI: 5.7152

Barplot representing computed co-occurrences of the lemma AQUA "water" in the *Patrologia Latina* corpus (data fetched from an R session exposed as an OpenCPU API; the chart generated with the *d3.js* library).

The external resources are exploited through their public APIs. This is also the case of the locally hosted services. Yet, their role is by no means limited to only exposing data, since they also serve to enrich, compute on and prepare data for subsequent display. For instance, an R session is exposed to the web application through the OpenCPU API (Ooms, 2014) and permits computation on corpusand lexicography resources: the rcqp package (Desgraupes and Loiseau, 2012) is used to connect to the CQP engine (Hardie, 2012), A. Guerreau's scripts for lexical statistics allow to find co-occurrences of the lemma in the corpus, and the wordspace package (Evert, 2014) is employed to calculate word similarities based on word's distributional properties.

In spite of the relative variety of technologies and formats used, the *medialatinitas.eu is* planned to be developed in a modular way, so the users could contribute their widgets as R or JavaScript code snippets responsible for self-contained functionalities.

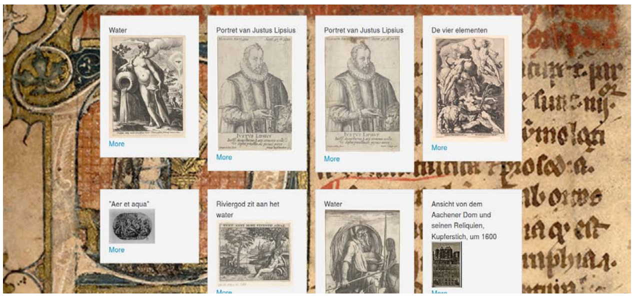

Water
More
Portret van Justus Lipsius
More
Portret van Justus Lipsius
More
De vier elementen
More
"Aer et aqua"
More
Riviergod zit aan het water
Water
Ansicht von dem Aachener Dom und seinen Reliquien, Kupferstich, um 1600

Image search (the lemma AQUA "water") in the Europeana repository.

Bibliography

Desgraupes, B. and Loiseau, S. (2012). *rcqp: Interface to the Corpus Query Protocol*. http://CRAN.R-project.org/package=rcqp (accessed 2 March 2016).

Evert, S. (2014). Distributional Semantics in R with the wordspace Package. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin: Dublin City University/Association for Computational Linguistics, pp. 110–14.

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, **17**(3): 380–409.

Nowak, K. and Bon, B. (2015). *medialatinitas.eu*. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin. In I. Kosem et al. (Eds.). *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 152–69.

Ooms, J. (2014). The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns. ArXiv e-prints. http://arxiv.org/pdf/1406.4806v1.pdf (accessed 1 March 2016).

Time Series Analysis Enhances Authorship Attribution

Jeremi K. Ochab
jeremi.ochab@uj.edu.pl
Jagiellonian University, Poland

Introduction

Long-range correlations in texts – emerging even in dictionaries and allowing to differentiate genres (Montemurro and Pury, 2002) – prove that structures larger than necessitated by syntax exist. They might reflect organisation of literary works, and be one of authorial fingerprints.

Stylometry, however, have not exploited the information carried by memory longer than one clause apart – other than use of n-grams (e.g., Eder, 2011). Existing studies include investigating: sequences of (un)stressed syllables (Pawłowski, 1998; 1999); sentence lengths (Drożdż et al., 2016); transferring long-range correlations between letter and word sequences (Altmann et al., 2012).

To quantify correlations in a script, successive symbols can be treated as a time-series with symbolic values (Stanley, 1992) or numeric positions on a word frequency list (Montemurro and Pury, 2002; Ausloos, 2012), see Fig. 1. Below, I use information extracted from such time-series as features in machine learning (ML) methods to increase accuracy of authorship attribution (AA) in a benchmark literary corpus.

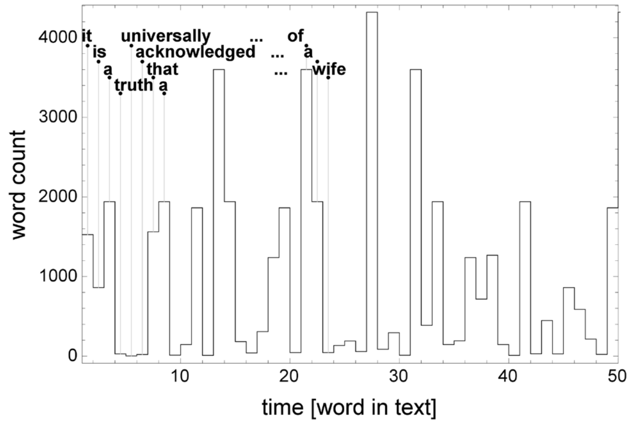


Figure 1 A book translated into a time-series: *x*-axis corresponds to the position of a word in a text; *y*-axis corresponds to the total number of times the word appears in the book

Methods

Complexity measures

For a series, defined as ranks of words at consecutive positions in a text, following measures are used (formatted as **Quantity**: *ML features*):

Power spectrum: *psLen*, *psExp*

Power spectrum *S(f)* of a series at a frequency *f* can be interpreted as the strength of correlation of the series with itself at word-to-word distances *1/f*. As Fig. 2 illustrates, it is described well by two parameters: the length *psLen* of the high-correlation plateau and the slope of its decay *psExp*.

Predictability: *pred*

As the name suggests, it measures how well can the next step in a series be predicted given previous steps (see definition: Stone, 1996).

Fano factor exponent: *fanoF*

Fano factor measures signal autocorrelation, especially in fractal processes (Thurner et al., 1997), as one takes increasingly bigger chunks of text – similarly to the slope of power spectrum or detrended fluctuation analysis (Grabska-Gradzińska et al., 2013).

Entropy rate of word variation: *entExp*, *entConst*

The entropy is maximal for equiprobable word occurences, and minimal when a single word is always used. As one reads a text, new words appear and the entropy grows, and saturates. *entExp* and *entConst* are characteristic time and a multiplicative constant of such a growth.

Static entropy: *entLocM*, *entLocSD*

For a window of a constant length moving across the whole text, the entropy fluctuates. Parameters *entLocM* and *entLocSD* are its mean and standard deviation.

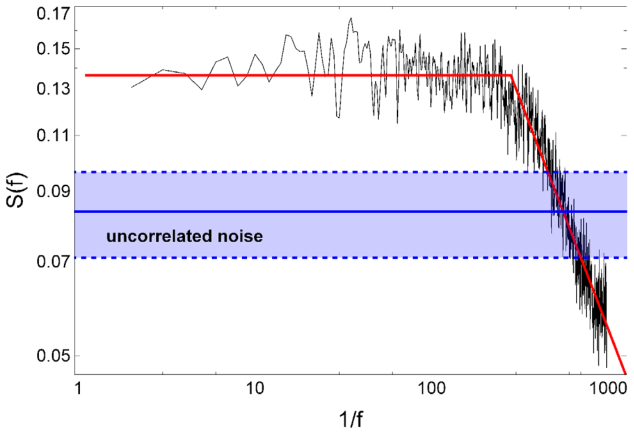


Figure 2 Power spectrum *S(f)* of the time-series as a function of word-to-word distance 1/f

Algorithm and parameters

AA was performed with the R package *stylo* (Eder et al., 2013) with settings: delta distance (Burrows, 2002), 1000-fold cross-validation, one book of each author in the training set. (None of the other ML methods (Stamatatos, 2009; Jockers and Witten, 2010) implemented in *stylo* did significantly better than Burrows's delta.)

Since on this corpus about 90 most frequent words (MFW) are needed for 100% accuracy, only the first ten were used as features, which left room for improvement. Having precomputed all the eight measures, they were appended to the feature list.

Data

A corpus (Rybicki, 2015) comprising 27 classic British 19th c. novels of 11 authors was used (see Fig. 3, where each leaf is a shorthand for a novel's *Author_Title*). The reason for choosing this corpus is that many AA algorithms have been tested on it, and they perform very well, not least thanks to its size.

Results

Authorship attribution

AA algorithms at best use 6-grams (Eder, 2011), whereas the correlations may reach hundreds of words, as demonstrated in Fig. 2. The results in Tables 1-2 show that the measures from Sec. 2.1 can aid ML. As a proof of concept, Fig. 3 shows a cluster analysis based exclusively on these complexity measures; although imperfect, it strongly indicates that the temporal characteristics contain traces of authorial style.

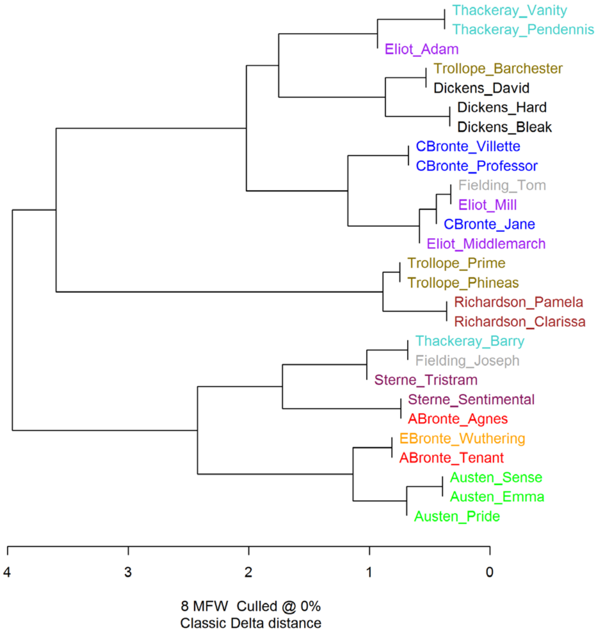


Figure 3 Cluster analysis of the corpus based only on the eight complexity measures

Informativeness of complexity measures

Surprisingly, *psLen* is not correlated with paragraph lengths (cf. Kosmidis et al., 2006). Its smallest values 280-300 come, intriguingly, from Austen and Anne and Emily Brontë, while the largest 370-390 from Dickens, Thackeray and Trollope.

Note that correlated features (see Tab. 3 for a summary) worsen performance and should be eliminated. Remaining parameters are expected to contain non-overlapping information. Further, PCA analysis showed that *psLen* and *entLocM* contain the most distinctive information. Tables 1-2 show that indeed these parameters most significantly aid ML.

Conclusions

This preliminary study shows that measures reflecting long-range word-to-word correlations carry authorial information and enhance stylometric ML methods. More complex features than words and n-grams are needed.

Acknowledgements

I thank Maciej Eder for insightful discussions. The research was funded by Grant No. DEC-2013/09/N/ST6/01419 of the National Science Centre of Poland.

Features	**Accuracy**±SD [%]	**Features**	**Accuracy**±SD [%]
10 MFW	87.14 ± 0.22	**psLen**	**91.14** ± 0.18
11 MFW	88.44 ± 0.22	**psExp**	**88.52** ± 0.21
88-89 MFW	100	**pred**	**87.61** ± 0.21
3 best c.m.	43.09 ± 0.35	fanoF	83.20 ± 0.24
		entExp	85.89 ± 0.21
		entConst	85.99 ± 0.21
		entLocM	**91.69** ± 0.16
		entLocSD	83.29 ± 0.24

Table 1 Left: Reference values **Right:** Clustering accuracy for 10 MFW+1 complexity measure as a feature (accuracy above 10-MFW reference level is in bold)

Features	**Accuracy**±SD [%]
psLen+entLocM	92.49 ± 0.16
psLen+entLocM+entExp	92.95 ± 0.15
psLen+entLocM+pred	92.53 ± 0.14

Table 2 Best accuracy obtained for 10 MFW plus two/three complexity measures

	psLen	psExp	entLocSD
psLen		− − −	+
psExp	− − −		
entLocSD	+		

Table 3 Positive (+) and negative (–) correlations between parameters from Sec. 2.1

Bibliography

Altmann, E. G., Cristadoro, G. and Esposti, M. D. (2012). On the origin of long-range correlations in texts, *Proceedings of the National Academy of Sciences*, **109**: 11582–87.

Ausloos, M. (2012). Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series, *Physical Review E*, **86**: 031108.

Burrows, J. F. (2002). "Delta": a measure of stylistic difference and a guide to likely authorship, *Literary and Linguistic Computing*, **17**: 267–87.

Drożdż, S., Oświęcimka, P., Kulig, A., Kwapień, J., Bazarnik, K., Grabska-Gradzińska, I., Rybicki, J. and Stanuszek, M. (2016). Quantifying origin and character of long-range correlations in narrative texts, *Information Sciences* **331**: 32-44.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint, *Studies in Polish Linguistics*, pp. 99–114.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts.* University of Nebraska-Lincoln, NE, pp. 487–89.

Grabska-Gradzińska, I., Kulig, A., Kwapień, J., Oświęcimka, P. and Drożdż, S. (2013). Multifractal analysis of sentence lengths in English literary texts, *AWERProcedia Information Technology and Computer Science* **3**: 1700-06.

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution, *Literary and Linguistic Computing*, **25**: 15–223.

Kosmidis, K., Kalampokis, A. and Argyrakis, P. (2006). Language time series analysis, *Physica A: Statistical Mechanics and its Applications*, **370**: 808–16.

Montemurro, M. A. and Pury, P. A. (2002). *Long-range fractal correlations in literary corpora, Fractals* **10**: 451–61.

Pawłowski, A. (1998). Séries temporelles en linguistique. avec application a lattribution de textes: Romain Gary et Émile Ajar. *Travaux de linguistique quantitative*, Vol. **62**, Honoré Champion, Paris, Geneve: Champion-Slatkine.

Pawłowski, A. (2011). Language in the line vs. language in the mass: On the efficiency of sequential modeling in the analysis of rhythm, *Journal of Quantitative Linguistics* **6**: 70–77.

Peng, C., Buldyrev, S., Goldberger, A., Havlin, S., Sciortino, F., Simons, M. and Stanley, H. (1992). Long-range correlations in nucleotide sequences, *Nature*, **356**: 168–70.

Rybicki, J. (2015). *A short collection of British fiction.*https://sites.google.com/site/computationalstylistics/corpora (accessed 23 February 2016).

Stamatatos, E. (2009). A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, **60**: 538–56.

Stone, J. V. (1996). Perceptually salient visual parameters using spatiotemporal smoothness constraints, *Neural Computation*, **8**: 1463–92.

Thurner, S., Lowen, S. B., Feurstein, M. C., Heneghan, C., Feichtinger, H. G. and Teich, M. C. (1997). Analysis, synthesis, and estimation of fractal-rate stochastic point processes, *Fractals*, **5**: 565–95.

Adding Flexibility to Large-Scale Text Visualization with HathiTrust+Bookworm

Peter Organisciak
organis2@illinois.edu
University of Illinois, United States of America

Sayan Bhattacharyya
sayan@illinois.edu
University of Illinois, United States of America

Loretta Auvil
lauvil@illinois.edu
University of Illinois, United States of America

Leena Unnikrishnan
unnikrishnan.leena@gmail.com
Indiana University, United States of America

Benjamin Schmidt
b.schmidt@neu.edu
Northeastern University, Massachusetts, United States of America

Muhammad Saad Shamim
sa501428@gmail.com
Baylor College of Medicine, United States of America

Robert McDonald
rhmcdona@indiana.edu
Indiana University, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois, United States of America

Erez Lieberman Aiden
erez@erez.com
Baylor College of Medicine, United States of America

The HathiTrust holds one of the largest collections of digitized published work in the world. With nearly 14 million volumes ("Currently Digitized", HathiTrust.org), it is a challenge to create mechanisms to understand and interpret a collection of this size. The HathiTrust+Bookworm (HT+BW) project presents ways to behold that textual content through interactive visualization. In this poster, we present new work from HT+BW in applying visualization as an complementary, rather than center, tool for supporting better comprehension of large collections. Whereas HT+BW has previously been used in standalone contexts with pre-determined metadata, we present work in two new areas: 1) allowing scholars to analyze custom personal collections from within the larger corpus; and 2) use of HT+BW as a supplement to other uses of the HathiTrust Research Center.

Context

The vanguard of big text data was the realization that we had more books than a person can read in a lifetime (Crane, 2006). At the scale of current collections, such as that of the HathiTrust, it would take many years to read only the *titles* of all the digitized works. Such a scale presents opportunities for new forms of inference about historical, cultural, and linguistic trends. However, because of the necessary abstractions away from qualitative reading to quantifiable features in these texts, the underlying biases of the collection are not always apparent to a scholar. HT+BW seeks to address such problems.

HathiTrust+Bookworm (HT+BW) is a project seeking to adapt the generalized analytic tool Bookworm (http://bookworm.culturomics.org) to the large scales and unique needs in the HathiTrust Research Center. The HathiTrust is a consortium of institutions collecting millions of digitized works, and its research center seeks to support scholars in the large-scale insights that such a collection can accommodate. HT+BW is built on the HathiTrust Research Center's Extracted Features Dataset (Capitanu et al., 2015), a publicly-released dataset of page-level counts for important features, such as text frequencies. The dataset contains nearly 8 terabytes of data for 1.8 billion pages across 4.8 million books, with plans to grow three-fold in 2016.

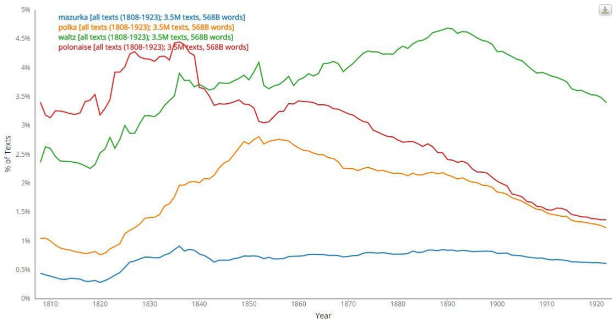
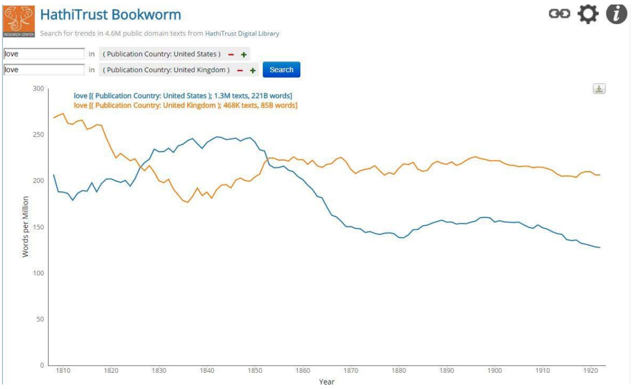


Figure 1: Comparing multiple topics over all texts



love in (Publication Country: United States)

love in (Publication Country: United Kingdom)

love [(Publication Country: United States); 1.3M texts, 221B words]

love [(Publication Country: United Kingdom); 468K texts, 85B words]

Figure 2: Comparing a single topic in multiple facets of the collections (US-published and UK-published books)

Approach

The strength of HT+BW is in detailed querying and faceting. In contrast to the Google Ngrams Viewer, for instance, scholars do not have to compare term usage across the entire corpus. Rather, they can peer within highly particular facets; for example, one can visualize longitudinal trends for the word "love" for only those books that were both published in the US and classified as literature (Figure 2). This ability becomes much more valuable when comparing across different facets, so that the usage trend of "love" can be compared between literature and general non-fiction works.

Our current work extends faceting to custom groupings of the collection — personal subsets called *worksets* — in the Hathitrust Research Center (Jett, 2015). Worksets can be visualized in Bookworm in a manner similar to visualizing any predetermined metadata-based facet. The affordances this allows include:

- generating descriptive statistics for a particular research domain, such as late 19th century best-sellers;
- creating two worksets, one of early-career and one of late-career authors, and comparing how theme words occur in either case; or
- using HT+BW over a collection derived from other scholarship, such as (Ted Underwood's fiction data).

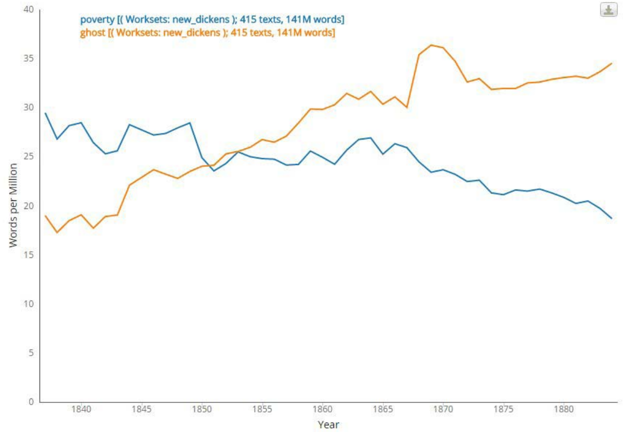


Figure 3: Comparing terms used in a 2000-volume scholar curated workset

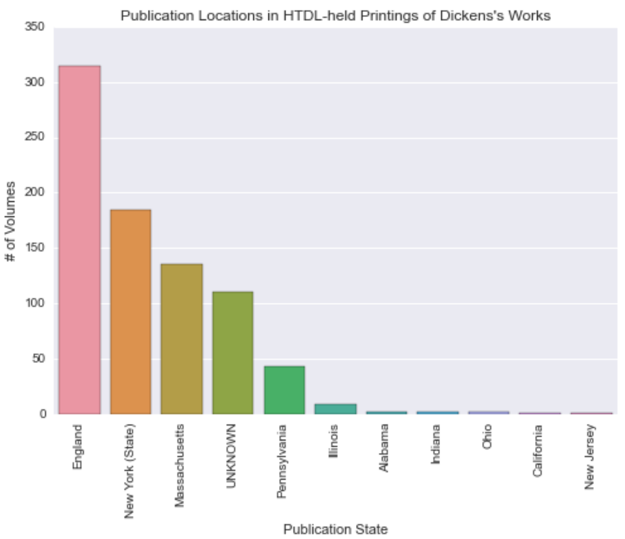


Figure 4: Comparing facet counts for "publication state" within a workset, using HT+BW public API.

In addition to workset access, HT+BW is being used for "widget" access to support other activities within the broader HathiTrust Research Center. Most importantly, this includes search. When searching for individual works, or building a workset, it is not immediately clear what relationship the relevant results have to the underlying contours of the dataset. For example, the widget-style visualization of a researcher's query for "Beijing" or "Istanbul" would reveal that those terms are biased toward 20th century work (as these places were earlier referred to by other names).

The primary development challenge of the HT+BW has been the technical hurdles inherent to the scale of the HathiTrust's collection. We have improved the ingest process for Bookworm to allow for lower level optimizations, a gain applicable to other projects' use of Bookworm. Query performance is robust at the scale of unigrams for 4.8 million volumes, though it remains to be seen how this will be scaled to larger n-grams and more volumes.

Conclusion

The HT+BW project is providing richer, more flexible access to the large holdings of the HathiTrust Digital Library, with a goal of supporting humanists in probing questions about historical, literary, or cultural trends in published literature.

Notes

HT+BW is supporting the development of Bookworm as a standalone, collection-agnostic tool. Our technical contributions are available at https://github.com/Bookworm-project and https://github.com/htrc. HT+BW is built upon publicly-available data from the HathiTrust Digital Library (https://sharc.hathitrust.org/features) and metadata (https://www.hathitrust.org/hathifiles).

HT+BW is supported by NEH Implementation Grant HK-50176-14. Any views, findings, conclusions, or recommendations do not necessarily reflect those of the National Endowment for the Humanities.

Bibliography

Currently Digitized. *Hathitrust Digital Library*. http://hathitrust.org/about.

Bookworm. *Culturomics*. http://bookworm.culturomics.org.

Boris, C., Underwood, T., Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C. and Downie, J. S. (2015). *Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain* Vol. **2**, [Dataset]. HathiTrust Research Center, doi:10.13012/j8td9v7m.

Crane, G. (2006). *What Do You Do with a Million Books?*, *D-Lib Magazine*, **12**(3). doi:10.1045/march2006-crane.

Jett, J. (2015). *Modeling Worksets in the HathiTrust Research Center*. CIRSS Technical Report WCSA0715. Champaign, IL: University of Illinois at Urbana-Champaign.

Cultivating Digital Humanities Biomes: A Collaborative Model

Thomas George Padilla
tpadilla@msu.edu
Michigan State University, United States of America

Kristen Mapes
kmapes@msu.edu
Michigan State University, United States of America

Brandon Locke
blocke@msu.edu
Michigan State University, United States of America

The Digital Humanities community spans disciplines, academic and professional ranks, and the full spectrum of student education. This community self-defines in myriad ways, which poses challenges to faculty, librarians, staff, and others working to design curriculum, support skill development and methodological fluency, allocate resources, and develop services in the Digital Humanities. The authors of this poster conducted a year long, multi-pronged, campus-wide study to gain a greater understanding of how individuals across a large research intensive university campus include Digital Humanities methods in their research and pedagogy, what their data needs and preferences are, where these parties need greater support, and where additional opportunities for partnerships within and across departments, colleges, and units.

Given the desire to provide a unified and complementary set of services, curriculum, resources, and collaborative possibilities, representatives from the libraries, one college, and a multi-disciplinary lab within another college at a research intensive university jointly designed and implemented a campus-wide survey and targeted, in depth interviews that equally reflected the Digital Humanities related missions of the aforementioned units. This poster demonstrates a collaborative model for assessing and cultivating a Digital Humanities community at a large research intensive university.

Michigan State University, home to MATRIX, H-Net, and WIDE, has been active in Digital Humanities and Social Sciences for more than two decades. Serendipitously, MSU Libraries, the College of Arts and Letters, and the Department of History each sought to build on this strength at the same time through a program of hiring. Amidst this constellation of Digital Humanities activity and inspired by the example of the University of Colorado, in which the university library conducted a Digital Humanities needs assessment, the authors resolved to conduct a data-driven analysis of the Digital Humanities environment on campus. To do so, the authors reached out to units known to the community in addition to disciplines less engaged with

Digital Humanities. The collaborative, cross-institutional nature of this needs assessment design is foundational as a model for other institutions to evaluate and enhance their own community and offerings.

The assessment model developed for this study adopted a holistic approach, focusing on current use and interest in methods and tools, programming languages, types of data used, data sources, and preferences for data access, preferences relating to collaboration, workshops and training sessions, and an inquiry into limitations and barriers to using digital and computational methods. The 20-minute long survey received 421 responses, including a strong turnout from key humanities and social science departments, including History, Anthropology, Linguistics and Languages, English, and Writing, Rhetoric, and American Cultures. Survey results were supplemented with ten hour-long interviews with faculty and graduate students about their use of digital methods and support on campus. Some respondents were selected by the survey convenors ahead of time, while others volunteered through the survey responses.

The results of the survey are being used at the unit level as well as university-wide at differing scales to face specific needs within units as well as to strategize campus-wide collaboration and capacity-building initiatives. The authors are using data at the unit level to augment support, services, and partnerships for pedagogy, research, data management, and skill development. Results are also being used to inform campuswide community building initiatives through outreach, strategic developments among units, and larger initiatives through the Office of Research.

This poster presents a model for developing a campus-wide Digital Humanities needs assessment intended to provide strategic information for multiple units on a large research intensive campus. By reviewing this model, other universities and colleges will come away with strategies for identifying what types of collaborations between colleges, departments, programs, labs, and libraries can improve DH pedagogy and research, what resources, activities, and services are needed to support those collaborations, as well as approaches for engaging those that are interested in the Digital Humanities but are not yet active participants. The poster presentation will include data derived from the needs assessment in addition to Digital Humanities community building solutions inspired by the data.

Curating Community: Building A Communications Strategy For The European Association For Digital Humanities

Eliza Papaki
e.papaki@dcu.gr
Digital Curation Unit, ATHENA R.C., Greece

James O'Sullivan
josullivan@psu.edu
Pennsylvania State University

Antonio Rojas
antonio.rojas@upf.edu
Universitat Pompeu Fabra

Summary

This poster will seek to illustrate the enhanced communications presence of the European Association for Digital Humanities (EADH) since 2014, and measure the impact of its action in building a community of geographically-dispersed members. It will outline the strategies undertaken by the EADH in furthering its communications initiative, for which a number of Communications Fellows have been activated on the organisation's communications policy in an effort to promote the work of digital scholars across the European region. Accounts of this activity will be further contextualised by theoretical discussion on the evolution of social media and Web-based communications across academia.

Introduction

This poster will seek to illustrate the enhanced communications presence of the European Association for Digital Humanities (EADH) since 2014, and measure the impact of its action in building a community of geographically-dispersed members. The poster will outline the strategies undertaken by the EADH in furthering its communications initiative over the last 24 months, during which a number of EADH Communications Fellows have been collaborating on the organisation's communications policy in an effort to promote, more broadly, the work of digital scholars across the European region. Accounts of this activity will be further contextualised by theoretical discussion on the evolution of social media and Web-based communications across academia.

Communications Fellows

As noted, in an effort to engage further with the international Digital Humanities community in respect to news, events, and opportunities, the EADH has motivated a diverse group of four junior scholars tasked with strengthening the Association's public interactions. The self-directed goals of the Fellows are as follows:

- Publish news, announcements and CfPs relevant to the European community of digital scholars
- Document projects undertaken in Europe in recent times and feature them in a slider to promote access and collaboration between members
- Ensure quality of language, as well as accuracy and detail of information, across all official EADH correspondence
- Increase community engagement through social media, particularly Facebook, Twitter, and LinkedIn
- Enhance the profile of relevant scholars and scholarship from across Europe's various Digital Humanities projects, centres, and initiatives
- Curate and disseminate information across a variety of European languages, enhancing the cultural diversity of the organisation's communications

Social Media

The social media revolution that has spread into the academy in recent years has shifted scholarly communications towards participatory technologies. Collectively, these social technologies are now dominating the ways in which users interact across the Web. In an effort to foster a network of researchers which embraces diversity, one must take advantage of the networks of exchange facilitated by these platforms. Initially, the focus of the fellows was oriented towards social media, Facebook, Twitter, and LinkedIn. These channels have proved essential in the collection and distribution of relevant information, and have contributed to the advancement of Europe's community of Digital Humanities scholars. Currently, the EADH engages with 2373 followers on Twitter, 491 users on Facebook, and 85 users on Linkedin. In the context of this networking activity, the EADH collaborates with associate organisations, AIUCD, Dhd, Nordic DH, and DH Benelux as well as with the Alliance of Digital Humanities Organizations (ADHO) to increase the synergies and the visibility of its members.

Discussion

This poster will illustrate the presence and performance of the EADH on social media, provoking discussion with attending researchers on their personalised use of social media and their future aspirations in regards to the employment of social media in promoting Digital Humanities research across regional and global contexts. Furthermore, the poster will present metrics on the success of this particular communications drive, as well as detail various strategic decisions, such as how to structure a Facebook group for a scholarly community, and the benefits and drawbacks of each.

Integrative 3D Recording Methods of Historic Architecture: Burg Hohenecken from Southwest Germany

Aaron Charles Pattee
acpattee30@gmail.com
University of Nebraska Lincoln, United States of America

Bernhard Höfle
hoefle@uni-heidelberg.de
Heidelberg University, Germany

Christian Seitz
christian.seitz@archeye.de
Heidelberg University, Germany

This study explores the integration of various methods including photogrammetry, laser-scanning, GIS, and textual analysis in order to create a more holistic understanding of a castle ruin. The case study is the medieval castle *Burg Hohenecken* in the city of *Kaiserslautern* in southwest Germany. The project is divided into two main components: the visual and the textual components. The objectives of the visual component are to merge a laser-scan and a photogrammetric model, thus combining the measuring strength of laser-scanning with the visual aesthetics of photogrammetry. Once digitized as a merged 3D-model, the castle can be virtually controlled and examined, providing an opportunity to potentially reconstruct the castle (using *Autodesk 3DsMax*). Hypothetical reconstructions from the various construction and expansion periods with assist in studying the castle's function throughout time. The textual component, consisting of 70 letters (1212-1560 A. D.), provides historical context concerning the castle, its inhabitants, and correspondence.

Laser-scanning and 3D photogrammetric technologies have been in rapid advancement for over 20 years beginning in the disciplines of geography and mathematics (Kersten et al., 2004); specifically in photogrammetry and geodesy (Vosselmann and Maas, 2010). Academics have developed a variety of techniques and devices in order to adapt the technologies for archaeological purposes (Remondino, 2014). The intricacies of archaeological sites have pushed the devices to their limits, allowing for the production of newer devices more suited for rough terrain and countless blind spots due to walls, vegetation, and elevations. The location of the castle atop a mountain adds to the difficulty of acquiring photos encompassing the entire castle in one take (Gonzo et al., 2004). As a result of these challenges, a terrestrial laser-scanner, an aerial drone with a mirror-less camera, and a terrestrial DSLR camera, were used.

The visual component consists of two part: the laser-scanning and the photogrammetry. The Riegl VZ-400 laser-scanner (with an attached DSLR camera) was provided by the Institute of Geography at Heidelberg University. We conducted 22 scans, manually linked via tie-points (placed within overlapping areas) in the proprietary software, *RiSCAN PRO*. This model was meshed in *Meshlab* using the *Ball-Pivot* method.

The photogrammetric data was collected using a Sony NEX-7 camera attached to an aerial drone built by the IWR (Interdisciplinary Center for Scientific Computing) at Heidelberg University, and a hand-held Nikon D3300 SLR camera (provided by the University of Nebraska-Lincoln). The eight rotor drone took 385 photos up to 100 meters above the castle in a predetermined course, capturing views not completely collected by the laser-scans. The remaining 1,351 photos were done terrestrially with the Nikon D3300 SLR camera. *Agisoft Photoscan Pro* was used to align the 1,736 photos, generate the dense cloud, the mesh, and the texture.

The textual component regards the information extracted from the letters, including names and inheritances of the *Hohenecken* lineage, physical locations within the surrounding environment, and construction details of the castle. This data builds a network of communication within a past landscape and pre-existing structures of the castle. The locations in the letters will be geo-referenced in GIS allowing the landscape to be visualized rather than imagined, thus reconstructing the environment of the past utilizing primary and secondary sources.

This project is a work in progress, with an anticipated completion by May 2016. An updated status can be viewed online (http://www.jamescoltrain.com/970.Aaron/) in which the letters and models have been uploaded. A virtual tour (using *Unity*) of the reconstructed 3D-model, based upon textual evidence will allow users to "walk" through and explore the castle. The purpose is to analyze the historic functions of the castle and its significance in the regional landscape, contributing to its preservation and enhancing the knowledge of the regional architecture.

Bibliography

Gonzo, L., El-Hakim, S., Girardi, S., Picard, M. and Whiting, E. (2004). Photo-Realistic 3D Reconstruction of Castles with Multiple-Sources Image-Based Techniques. *Proceedings of the ISPRS XXth Congress*. Istanbul, Turkey, pp. 120-25.

Kersten, T., Pardo, C. A. and Lindstaedt, M. (2004). 3D Acquisition, Modelling and Visualization of North German Castles by Digital Architectural Photogrammetry. *ISPRS WG V/2 Scene Modeling and Virtual Reality*.

Remondino, F. (Ed), (2014). *3D Recording and Modelling in Archaeology and Cultural Heritage*. Oxford: British Archaeological Reports.

Vosselmann, G. and Maas, H. G. (2010). *Airborne and Terrestrial Laser Scanning*. Dunbeath, Scotland: Whittles Publishing.

WebSty – an Open Web-based System for Exploring Stylometric Structures in Document Collections

Maciej Piasecki
maciej.piasecki@pwr.edu.pl
Wrocław University of Technology, Poland

Tomasz Walkowiak
tomasz.walkowiak@pwr.edu.pl
Wrocław University of Technology, Poland

Maciej Eder
maciejeder@gmail.com
[1] Institute of Polish Language, Polish Academy of Sciences;
[2] Pedagogical University of Kraków

Introduction

Computer-assisted text analysis is now witnessing the phenomenon of ever-growing computer power and, more importantly, an unprecedented aggregation of textual data. Certainly, it gives us an unique opportunity to see more than our predecessors, but at the same time it presents non-trivial challenges. To name but a few, these include information retrieval, data analysis, classification, genre recognition, sentiment analysis, and many others. It can be said that, after centuries of producing textual data and decades of digitisation them, the scholars now face another great challenge - that of beginning to make good use of this treasure.

Generally speaking, the problem of large amounts of textual data can be perceived from at least three different perspectives. Firstly, there is a need of asking new research questions that would take advantage of thousands of texts that can be compared. Secondly, one has to introduce and evaluate statistical techniques to deal with vast amounts of data. Thirdly, there is a need of new computational algorithms that would be able to handle enormous corpora, e.g. containing billions of tokens, in a reasonable amount of time. The present study addresses the third of the aforementioned issues.

Stylometric techniques are known for their high accuracy of text classification, but at the same time they are usually quite difficult to be used by, say, an average literary scholar. In this paper we present a general idea, followed by a fully functional prototype of an open stylometric system that facilitates its wide use with respect to two aspects: technical and research flexibility. The system relies on a server installation combined with a web-based user interface. This frees the user from the necessity of installing any additional software. Moreover, we plan to enlarge the set of standard stylometric features with style-markers referring to various levels of the natural language description and based on NLP methods.

Multi-aspectual Document Representation

Computing word frequencies is simple for English, but relatively complicated for highly inflected languages, e.g. Polish, with many word forms, resulting in data sparseness. Thus, it might be better first to map the inflected forms onto *lemmas* (i.e. basic morphological forms) with the help of a morpho-syntactic tagger, and next to calculate the lemma frequencies.

Most frequent words or lemmas as descriptive features proved to be useful in authorship attribution. However, for some text types or genres they do not provide sufficient information to tell the authors apart, e.g. see (Rygl, 2014). Moreover, in other types of classification tasks, where the goal is to trace signals of individual style, literary style or gender, it usually turns out that they appear on different levels of the linguistic structures. Thus, one needs to enhance text description.

In practice, every language tool introduces errors. However, if the error level is relatively small and the errors are not systematic (i.e. their distribution is not strongly biased), than the results of such a tool can be still valuable for stylometric analysis. Bearing this in mind, we have evaluated a number of language tools for Polish, and selected types of features to be implemented:

- length of: documents, paragraphs or sentences (a segmentation tool),
- morphological features
  - word forms or tokens and punctuation marks,
  - pseudo-suffixes (last several letters),
  - lemmas (from WCRFT2 morpho-syntactic tagger for Polish (Radziszewski, 2013))
- grammatical classes
  - 35 grammatical classes defined in the tagset (Szałkiewicz and Przepiórkowski, 2012) of the Polish National Corpus (Przepiórkowski et al., 2012), e.g. pseudo-past participle, non-past form, ad-adjectival adjective; recognised by WCRFT2,
  - parts of speech (by grouping grammatical classes),
  - grammatical categories, e.g. gender, number, person, etc.; (WCRFT2),
- sequences
  - lemma sequences (e.g. potential collocations),
  - sequences of grammatical classes (bigrams or trigrams - hints about the grammatical structures),
- semantic features
  - semantic *Proper Name classes* – recognised by a Named Entity Recogniser Liner2 (Marcinczuk, 2013),
  - *lexical meanings* – represented by synsets in plWordNet (the Polish wordnet); assigned by Word Sense Disambiguation tool WoSeDon (Kędzia, et al., 2015)
  - generalised lexical meanings – more general meanings from plWordNet, e.g. *an animal* instead of *a cheetah*,
  - formalised concepts from a formal ontology SUMO that plWordNet is mapped to,l
  - exicographic domains from wordnet.

Semantic features go beyond a typical stylometric description, but allow for crossing borders between the style and the content description.

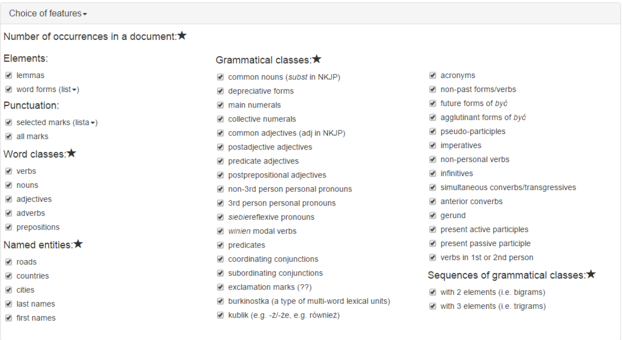
There are no features overtly describing the syntactic structure, as the available parsers for Polish express too high level of errors. The set of features can be further expanded by user-defined patterns expressed in the WCCL language (Radziszewski et al., 2011) that can be used to recognise binary relations between words and their combinations.

WebSty allows for testing the performance of the aforementioned features in different stylometric tasks, several case-studies will be presented on a set of Polish novels.

Processing and Results

The proposed system follows a typical stylometric workflow which was adapted to the chosen language tools and other components of the architecture (see Section 4).

1. Uploading a corpus of documents together with meta-data in CMDI format (Broeder et al., 2012) from the CLARIN infrastructure.
2. Choosing the features for the description of documents – done by the users (see Fig. 1).
3. Setting up the parameters for processing (users).
4. Pre-processing texts with the help of language tools.
5. Extracting the features from the pre-processed texts.
6. Calculating feature values.
7. Filtering and/or transforming the original feature values.
8. Clustering the feature vectors representing documents.
9. Extracting features that are characteristic for different clusters.
10. Presenting the results: visualisation or export of data.



Choice of features▾

Number of occurrences in a document:★

Elements:

☑ lemmas
☑ word forms (list▾)

Punctuation:

☑ selected marks (lista▾)
☑ all marks

Word classes:★

☑ verbs
☑ nouns
☑ adjectives
☑ adverbs
☑ prepositions

Named entities:★

☑ roads
☑ countries
☑ cities
☑ last names
☑ first names

Grammatical classes:★

☑ common nouns (*subst* in NKJP)
☑ depreciative forms
☑ main numerals
☑ collective numerals
☑ common adjectives (adj in NKJP)
☑ postadjective adjectives
☑ predicate adjectives
☑ postprepositional adjectives
☑ non-3rd person personal pronouns
☑ 3rd person personal pronouns
☑ *siebie*reflexive pronouns
☑ *winien* modal verbs
☑ predicates
☑ coordinating conjunctions
☑ subordinating conjunctions
☑ exclamation marks (??)
☑ burkinostka (a type of multi-word lexical units)
☑ kublik (e.g. -ż/-że, e.g. również)
☑ acronyms
☑ non-past forms/verbs
☑ future forms of *być*
☑ agglutinant forms of *być*
☑ pseudo-participles
☑ imperatives
☑ non-personal verbs
☑ infinitives
☑ simultaneous converbs/transgressives
☑ anterior converbs
☑ gerund
☑ present active participles
☑ present passive participle
☑ verbs in 1st or 2nd person

Sequences of grammatical classes:★

☑ with 2 elements (i.e. bigrams)
☑ with 3 elements (i.e. trigrams)

Fig.1 Choice of features GUI

The step 5 can be performed as: simple counting of words or lemmas, processing and counting annotations matching some patterns or running patterns for every position in a document. This is why a dedicated feature extraction module comes into stage, namely *Fextor* (Broda et al., 2013).

Filtering and transformation functions can be built into the clustering packages (see below) or implemented by the dedicated systems, e.g. *SuperMatrix* system (Broda and Piasecki, 2013).

The WebSty architecture can be relatively easy adapted to the use of any clustering package. The prototype is integrated with *Stylo* – an elaborated clustering package dedicated to stylometric analysis, and *Cluto* – a general clustering package (Zhao and Karypis, 2005). Stylo offers very good visualisation functionality, while *Cluto* delivers richer possibilities for formal analysis of the generated clusters. The obtained results are displayed by the web browser (see Fig. 2). Users can also download log files with formalised description of the clustering results.

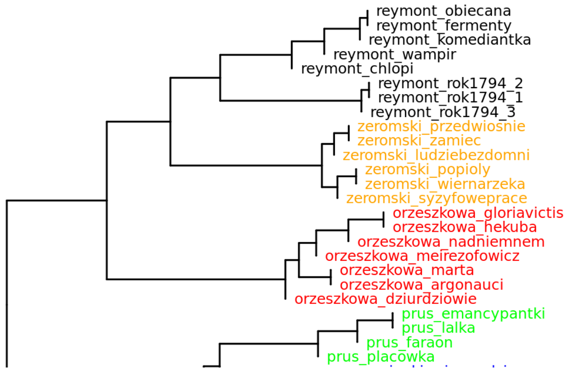


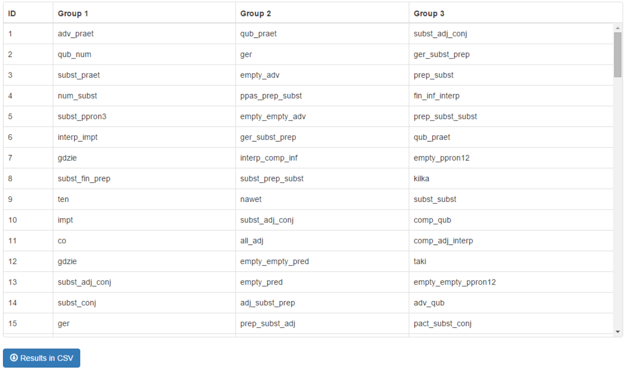
Fig.2 Stylometric results

Moreover, features that are characteristic for the description of individual clusters or differentiation between clusters can be identified. Ten different functions (implemented in Weka[1] (Witten et al., 2011) and SciPy packages[2]), based on mathematical statistics and information theory, are offered. The ranking lists of features are presented on the screen for interactive browsing (Fig. 3) and can be downloaded.

The system has a web-based user interface that allows for uploading input documents from a local machine or from a public repository and enables selecting a feature set, as well as options for a clustering algorithm.

Technical Architecture

Application of language tools is inevitably more complex than calculating statistics on the level of words or letters. Moreover, processing of Polish is mostly far more complex than English (in terms of the processing time and memory used). Thus, higher computing power and bigger resources are required. In order to cope with this, the entire analysis in WebSty is performed on a computing cluster. Users do not need to install any software - which might be non-trivial particularly in the case of the language tools. The system processes the documents using a parallel and distributed architecture (Walkowiak, 2015).



ID	Group 1	Group 2	Group 3
1	adv_praet	qub_praet	subst_adj_conj
2	qub_num	ger	ger_subst_prep
3	subst_praet	empty_adv	prep_subst
4	num_subst	ppas_prep_subst	fin_inf_interp
5	subst_ppron3	empty_empty_adv	prep_subst_subst
6	interp_impt	ger_subst_prep	qub_praet
7	gdzie	interp_comp_inf	empty_ppron12
8	subst_fin_prep	subst_prep_subst	kilka
9	ten	nawet	subst_subst
10	impt	subst_adj_conj	comp_qub
11	co	all_adj	comp_adj_interp
12	gdzie	empty_empty_pred	taki
13	subst_adj_conj	empty_pred	empty_empty_ppron12
14	subst_conj	adj_subst_prep	adv_qub
15	ger	prep_subst_adj	pact_subst_conj

Fig.3 Browsing significant features identified for the extracted clusters

The workflow is as follows. Input documents are processed in parallel. The uploaded documents are first converted to an uniform text format. Next, each text is analysed by a part-of-speech tagger (we use WCRFT2 for Polish (Radziszewski, 2013)) and then it is piped to a name entity recognizer - Liner2 (Marcińczuk et al., 2013) in our case. After the annotation phase is completed for all texts, the feature extraction module comes into stage, i.e. Fextor (Broda et al., 2013). Clustering tools requires input data in different formats: sparse or dense matrices, text (ARRF, Cluto format) or binary files (R objects, Python objects). Thus data received from the feature extraction for each input file has to be unified and converted. The extracted raw features can be filtered or transformed by a range of methods inside the clustering packages or in a system for distributional semantics called SuperMatrix (Broda and Piasecki, 2013). Finally, the R package Stylo (Eder et al., 2013) or a text clustering tool called Cluto (Zhao and Karypis, 2005) are run to perform exploratory analysis, e.g. multidimensional scaling.

To prevent the system from overloading and long response time the input data size is limited to 20 files. However, large text collections can be processed, if they are deposited in the dSpace repository.[3] All corpora in dSpace can be annotated stored for further processing. Therefore, it is only left to run feature extraction and clustering tools inside WebSty.[4]

Conclusion and future plans

The paper presented opened, web-based system for exploring stylometric structures in Polish document collections. The web based interface and the lack of the technical requirements facilitates the application of text clustering methods beyond the typical tasks of the stylometry, e.g. analysis of types of blogs (Maryl, 2012), recognition of the corpus internal structure, analysis of the subgroups and subcultures, etc.

The system is currently focused on processing Polish. However, as the feature representation is quite language independent, we plan to add converters for for other languages.

Bibliography

Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R. and Wardyński, A. (2013). Fextor: A feature extraction framework for natural language processing: A case study in word sense disambiguation, relation recognition and anaphora resolution. In Przepiórkowski, A., Piasecki, M., Jassem, K., Fuglewicz, P. (Eds.), Computational Linguistics: Applications, Series: *Studies in Computational Intelligence*, Vol. **458**, Springer Berlin Heidelberg, pp. 41-62.

Broda, B. and Piasecki, M. (2013). Parallel, Massive Processing in SuperMatrix – a General Tool for Distributional Semantic Analysis of Corpora. *International Journal of Data Mining, Modelling and Management*, **5**(1): 1–19.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T. and Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), pp. 1387-90.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: *Digital Humanities 2013: Conference Abstracts*. University of Nebraska-Lincoln, NE, pp. 487-89.

Kędzia, P., Piasecki, M. and Orlińska, M. J. (2015). Word Sense Disambiguation Based on Large Scale Polish CLARIN Heterogeneous Lexical Resources. *Cognitive Studies | Études cognitives*, **15**: 269-92.

Marcińczuk, M., Kocoń, J. and Janicki, M. (2013). Liner2 - A Customizable Framework for Proper Names Recognition for Polish. In Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., Niezgódka, M., Intelligent Tools for Building a Scientific Information Platform, Series: *Studies in Computational Intelligence,* Springer: Berlin Heidelberg, **467**: 231-53.

Maryl, M. (2012). Kim jest pisarz (w internecie?). *Teksty Drugie*, **6**: 77-100.

Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (Eds.),(2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.

Radziszewski, A. (2013). A tiered CRF tagger for Polish, In Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., Niezgódka, M., Intelligent Tools for Building a Scientific Information Platform, Series: *Studies in Computational Intelligence*. Springer Berlin Heidelberg, **467**: 215-30.

Radziszewski, A., Wardyński, A., and Śniatowski, T. (2011). *WCCL: A morpho-syntactic feature toolkit*. In Habernal, I. and Matoušek, V. (Eds.), Text, *Speech and Dialogue*, Plzen, Springer: Berlin Heidelberg, LNAI 6836, pp. 434–41.

Rygl, J. (2014) Automatic Adaptation of Author's Stylometric Features to Document Types, In Sojka, P., Horák, A., Kopeček, I. and Pala, K. (Eds.), *Proceedings of 17th International Conference TSD 2014*. Brno, Czech Republic, LNCS 8655, Springer: Berlin Heidelberg, pp. 53-61.

Szałkiewicz, Ł. and Przepiórkowski, A. (2012). *Anotacja morfoskładniowa*. In Przepiórkowski et al., pp. 59-96.

Walkowiak, T. (2015). Web based engine for processing and clustering of Polish texts. In Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J., *Proceedings of the Tenth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*, Brunów, Poland, Series: Advances in Intelligent Systems and Computing Springer, Springer Berlin Heidelberg, pp. 515-22.

Witten, I. H., Frank, E. and Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques, Third Edition. *Series in Data Management Systems*, Morgan Kaufmann.

Zhao, Y. and Karypis, G. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, **10**(2): 141-68.

Notes

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.scipy.org/

[3] https://clarin-pl.eu/dspace/

[4] http://ws.clarin-pl.eu/demo/cluto2.html

A new approach to libraries in the Digital Humanities: the case of "Fonte-Gaia"

Elena Pierazzo
elena.pierazzo@u-grenoble3.fr
Université Grenoble Alpes, France

Filippo Fonio
filippo.fonio@u-grenoble3.fr
Université Grenoble Alpes, France

Claire Mouraby
claire.mouraby@upmf-grenoble.fr
SID2 Grenoble, Université Grenoble Alpes, France

One of the much discussed topic of the digital humanities as an infrastructure concerns the role of libraries, which are not only the place where knowledge has been preserved for centuries (if not millennia), but also they represent the ideal partner for researchers and research centers, since they preserve most of the objects studied by scholars. Libraries have been identified as major actors for the preservation of digital artifacts (Van Zundert and Boot, 2012; Pierazzo, 2015), and many have engaged in such activities (Stanford Library, Brandeis University, British Library, for instance). In the same time the creation and the provision of digital libraries has now become almost normal, with most research libraries digitizing at least some part of their beholding and putting them online.

Although preservation and digitization are indeed of a fundamental importance, it is arguable that the mission of a library in the age of the digital can be extended even further. At the University of Grenoble a new and ambitious project has started. The idea is to transform the library of Italian Studies (which has a privileged status within the French library system) into a cultural hub, thanks to the tools and approaches offered by the Digital Humanities. The project is called "Fonte Gaia" (happy source), a name that plays with the double meaning of the Italian word "fonte", which means "fountain" (like the fountain in Siena from which the project is named) and "source" as in the sources of knowledge (the books); the "happy" refers to the conviviality of sharing such knowledge. The project is the expression of a larger consortium of libraries beholding considerable Italian books collections (Paris, Padua, Bologna, Rome, for now), and includes several components, surrounding the provision of a digital library, offering access to books digitized in Grenoble and in the other partner libraries. The digitized books will be offered in correlation with a series of tools and facilities for users: to annotate them; to create digital editions based on the same; to create reading paths and exercises. A funded PhD student is now investigating the functionalities and the ergonomics of an interface offering so many options to the users. The project has a strong pedagogical component: from the involvement of BAs and MAs students in the workflow of production of the eBooks, to PhD studentships, to the hosting of interns from Italy and France, to the organization of an French-Italian summer school for digital editing (starting summer 2016) and workshops; on that context a partnership with the ITN DiXit has been put in place.

In addition to these initiatives, two periodical publications are being proposed: first an informal un-journal (inspired by the format of un-conferences), where a group of early career researchers animate a polyphonic blog on topics rotating around Italian Studies and the time of digital; and second a more formal, peered-review scholarly journal which will be focused on libraries, Italian Studies and digital humanities. While this latter is still at the planning stage, the former has already started in April 2015 with great enthusiasm and success, with encouraging numbers of visits and good feedback from scholars, librarians and the Open Access environments in Italy and France. In fact "Fonte Gaia" adopt a strict golden OA policy for all the contents that are generated under its label, as well as for the content that are harvested; "Fonte Gaia" actively engages in the diffusion of OA policies and initiatives.

On the research side, "Fonte Gaia" incubates and fuels several small and medium research projects, providing support to early career researchers and other scholars in the mounting of digital humanities projects. At moment we have just started a project on metadata standards to open up our content to a linked data approach; two other new projects are in the phase of elaboration, one centered on the private archive of one of the most prominent translator of Italian literature in French between the end of the Nineteenth and the beginning of the Twentieth century, and the second on the census and digital cataloguing of Italian manuscripts possessed by the libraries of Grenoble. Connected to this latter, a module for virtual exhibition will be developed.

One of the aspects that make "Fonte Gaia" standing out from other similar initiatives is a partnership on equal basis between research and teaching centers and libraries. The different sets of complementary competences and experiences are at the base of the creation of a center of cultural animation, as well as preservation and production in a digital context. The pedagogical vocation shared by teachers and librarians are also shaping a unique learning environment with a strong hands-on component coupled with theoretical and ethical teaching. This ethos reflects also on the research and cultural heritage components of the project, making of "Fonte Gaia" a pioneering experiment of bringing together people, expertise and innovation.

This type of partnership is a genuine revolution in the French context, where traditionally the relationship between librarians and scholars is more a matter of a polite sharing of spaces,often verging on indifference, rather than one of active cooperation. In fact, since its inception this equal governance of "Fonte Gaia" has produced astonishment (luckily sometimes curiosity as well) among the members of the two categories as it challenges their tradition, their missions and their working methods. We take this as proof that we have to persevere.

Bibliography

Van Zundert, J. and Boot, P. (2012). The Digital Edition 2.0 and the Digital Library: Services, not Resources, *Bibliothek and Wissenschaft*, **44**: 141-52.

Pierazzo, E. (2015). Digital Scholarly Editing. *Theories, Methods and Models* (Aldershot: Ashgate).

Fonte Gaia Blog. http://fontegaia.hypotheses.org.

Agréger le passé en ligne: Euchronie, le passé ici et maintenant!

Sébastien Poublanc
sebastien.poublanc@gmail.com
Euchronie, France

Rémy Besson
remybesson@gmail.com
Euchronie, France

Euchronie est un projet collaboratif francophone s'inscrivant dans le domaine des humanités numériques. Partant du constat qu'une abondante production de contenu sur le passé est autopubliée sur le web francophone, le projet vise créer une base de données en centralisant, collectant, indexant et hiérarchisant dès leur parution cette multiplicité de contenus.[1] Contemporains de l'essor de la blogosphère, puis des humanités numériques, de l'autoproduction dans le domaine musical tout autant que de la diffusion de la photographie amateur en ligne, ces contenus ont pour double spécificité d'émaner de chercheur.e.s et d'être autopubliés sur le web.[2]

La base de données ainsi créée prend la forme d'un site web hébergé par Huma-Num, la très grande infrastructure de recherche numérique pour les SHS. Pour parvenir à créer rapidement un corpus d'étude, la collecte de données est facilitée par l'utilisation du plug-in PressForward pour WordPress. Libre et gratuit, aisé à prendre en main par des chercheurs néophytes en humanités numériques, il autorise la recopie partielle de pages produites sur différents sites web et réseaux sociaux, mais aussi de vidéos et de fichiers audio. Dans le même temps, l'agrégation des contenus s'accompagne d'une éditorialisation fondée sur la hiérarchisation des données : nous proposons un nouvel agencement basé sur l'ajout de métadonnées. Strictement normées, ces dernières permettent une catégorisation fine (chronologie, thématique et aire géographique) qui constitue l'une des principales valeurs ajoutées du projet.

Pour développer le contenu de la base de données, chaque membre du comité éditorial d'Euchronie est chargé de la veille d'un ensemble de sites défini en fonction de son domaine de spécialité. Pour chaque actualité publiée, le responsable éditorial s'assure qu'il s'agit bien de l'expression d'un point de vue, sans prise de position partisane : il s'agit de la dimension qualitative de la sélection. D'autre part, le responsable éditorial n'est pas amené à faire une sélection sur des critères plus restrictifs tels que l'intérêt historiographique du contenu ou une quelconque adéquation avec une ligne éditoriale que la plateforme souhaiterait défendre. Au contraire, la volonté à la base de la création de cette plateforme est d'être un lieu de rencontre entre différents types de discours sur le passé. À ce titre, l'aspect collaboratif du projet est primordial : dans la mesure où l'indexation est avant tout pensée comme une interaction sociale, les usagers de la plateforme sont aussi invités à proposer l'agrégation de billets ou sites qu'ils et elles ont identifiés. Enfin, un billet de synthèse est publié chaque semaine sur le blog du projet. Dans celui-ci, chaque membre du comité éditorial y présente brièvement les trois meilleurs contenus de la semaine.

Le site web permet ainsi de rendre compte d'une écriture de l'histoire de moins en moins logocentrée et de plus en plus intermédiale, c'est-à-dire écrite et visuelle, parlée et sonore, interactive et collaborative. Euchronie est ainsi un outil créé par et pour des spécialistes, tout autant qu'un point d'entrée pour les passionnés du passé.

En phase de développement, le projet dispose d'un financement assuré par plusieurs laboratoires de recherche en histoire et information-communication. Au 1er novembre 2015, son comité éditorial se compose de 19 membres séniors et juniors en France, Suisse et Canada, tandis que la première version de son corpus agrège 100 références, tant au format textuel que vidéo et audio. La mise en ligne du site est prévu pour l'année 2016 et sera donc visualisable lors du colloque. Un blog Hypotheses regroupe l'ensemble des informations relatives au projet.

La création de notre poster doit permettre de présenter l'outil et d'échanger autour des problématiques liées à la construction d'un corpus dans une perspective de Digital Public History. Nous désirons notamment rencontre les développeurs de PressForward pour partager nos retours et comparer notre projet avec Digital History Now.

Notes

[1] Notes historiographiques, synthèses méthodologiques, premières versions d'articles en cours de rédaction, hypothèses d'un chapitre de thèse encore à écrire, analyses d'archives en ligne, tutoriels concernant de nouveaux outils, expression de doutes épistémologiques, recensions d'ouvrages récents ou de classiques, analyses de productions filmiques ou multimédias, prises de position dans l'espace public, micro-essais d'ego histoire, vidéos, podcasts audio…

[2] Doctorant.e.s aussi bien que professeur.e.s d'université, enseignant.e.s du secondaire, amateurs et amatrices éclairé.e.s faiant usage des méthodes de la discipline.

Commens Digital Companion to Charles S. Peirce

João Queiroz
queirozj@gmail.com
Federal University of Juiz de Fora

Mats Bergman
mats.bergman@helsinki.fi
University of Helsinki

Sami Paavola
sami.paavola@helsinki.fi
University of Helsinki

Background

The Commens Digital Companion (http://www.commens.org) was born in 2012-13, when the Helsinki-based Commens site, designed by Mats Bergman and Sami Paavola, merged with the Brazilian Digital Encyclopedia of Charles S. Peirce, founded by João Queiroz. The new international Companion brought together two established on-line resources of Peirce research – the Commens Dictionary and the Digital Encyclopedia – and added new tools, such as the news service and the bibliography.

The Original Commens

Commens was originally developed as a Finnish Peirce studies website by Mats Bergman and Sami Paavola. It functioned as the virtual home for the Helsinki Metaphysical Club, and offered resources for students and researchers in both Finnish and English. The original Commens was opened in February, 2001. In 2003, Bergman and Paavola introduced the Commens Dictionary of Peirce's Terms, a dictionary of Peircean terminology built from original Peirce quotes. The Dictionary proved to be the most successful part of Commens; together with the Digital Encyclopedia, the Dictionary forms the backbone of the new Digital Companion. More than 700 quotes were transferred from the old Dictionary to the new platform.

Why "Commens"? Why a "Digital Companion"?

Peirce introduced the concept of the "commens" or "commind" in his correspondence with Victoria Lady Welby in 1906. Although a rare and partly obscure neologism, the term has proven to be both suggestive and useful. When Bergman and Paavola created the original Commens, their goal was a web site that would serve as an enabler of scholarly communication as well as a free resource for researchers and students - and hence "Commens", a term that suggests community and sharing of knowledge. The new version of Commens has also been developed in this spirit; it moves us several steps closer to the original aim by providing new possibilities for user input and collaboration.

In addition to evoking the original Digital Encyclopedia created by João Queiroz, the term "Digital Companion" refers to the emphasis on the new forms of publication that Commens provides. Unlike traditional, printed companion volumes, Commens grows and evolves through user input.

Primary Contents

News

Commens collects news pertaining to Peirce studies and closely related fields. The news are divided into categories, such as "academic meetings", "calls for papers", and "publications". An archive of all news items is provided on the news page. In addition, news with a set date, such as meetings and deadlines, are listed in the "Upcoming Events and Deadlines" sidebar block on the news page and the front page as well as the Commens calendar. All registered members may add news to Commens, as long as the entries relate to Peirce studies in some manner. All news items are supervised by the site editors, who have the right to remove any inappropriate entries.

Dictionary

The Commens Dictionary consists of original quotations, in which Peirce defines or characterises his technical terms. This collection of quotes does not lay claim to completeness; the dictionary aims to offer representative samples of Peirce's terminological definitions. It is intended to serve as an aid for researchers and students. All quotes are associated with sources (such as manuscripts or journal articles) added to the bibliography. Only editors and contributors may add and edit the contents; but all registered users may add comments (such as suggestions for new quotes and corrections). The comments are monitored by the site editors, who have the right to remove any inappropriate comments.

Encyclopedia

The Commens Encyclopedia contains original research articles related to different aspects of the philosophy and life of Peirce. The articles are peer reviewd according to academic standards.

Bibliography

The Commens Bibliography is an open-ended biographical database that contains both primary and secondary Peirce research sources. The bibliography provides the source references for the Commens Dictionary. All registered users can add records to the bibliography as long as the entries relate to Peirce studies in some manner. The bibliography is supervised by the site editors, who have the right to remove any inappropriate entries.

Bibliography

Bergman, M., Paavola, S. and Queiroz, J. (2013). *Commens Digital Companion of C. S. Peirce.* Available at http://www.commens.org/

Excerpta Constantiniana: From Palimpsest to a Digital Edition of a Medieval Encyclopaedia

Dariya Rafiyenko
dariya.rafiyenko@uni-leipzig.de
University of Leipzig, Germany

The aim of the poster is to present a digital edition of the *Excerpta Constantiniana* (*Excerpta* further on), a Byzantine encyclopaedia written in Ancient Greek in Constantinople in the 10th century (edited by de Boor, 1903, 1905; Büttner-Wobst, 1906; Boissevain, 1906; Roos, 1910).

The underlying research project belongs to the field of Classical and Byzantine Philology and is devoted to the edition of an important historiographical source. The goal of this poster is interdisciplinary: it aims to define the role of the editor and the concept of the presentation of a historical source in the digital environment.

The *Excerpta* was planned as a large-scale encyclopaedia in multiple volumes. It consists of several thousand separate extracts (excerpts) that were taken from about three dozens of Ancient Greek and Byzantine historiographical works. The extant parts of the *Excerpta* contain about 560 000 words (it is believed that almost tenfold is lost). We possess two original manuscripts of the *Excerpta* (each of the two for a different volume of *Excerpta*) that distinguish themselves through a remarkable *mise-en-page*: for the purpose of navigation in the content several hundred notes and pictograms have been placed on the margins of the manuscript (see image 2).

The digital edition of the whole work is in the preparatory phase; a section from the *Excerpta* have been edited exemplary, to be more precise, 24 pages (about 9 000 words) from the original manuscript *Vaticanus graecus* 73. The manuscript itself is a palimpsest. The text of the *Excerpta* has been washed off in the 14th century and overwritten, leading to a limited legibility of the original text (see image 1). Standard solution in such a case would be to publish the facsimile of the manuscript and to prepare a historical-critical edition of the text. However, the publication of such a damaged text proved itself to be of little value for the reader not versatile in deciphering palimpsests. Moreover, the traditional layout of a historical-critical edition is not suitable for displaying the *Excerpta*. The biggest challenge here is a faithful reproduction of the notes and pictograms on the margins that are important for the understanding of the text.

The *mise-en-page* of the *Excerpta* was pivotal for the development of the concept of this digital edition. A *pluralistic* approach to the text was taken as a basis. The goal was to ensure the presentation of the text in multiple *views* (displaying perspectives), most important of which were the *document-focused* and the *text-focused* perspectives (cf. Pierazzo and Stroke, 2011; Rehbein and Gabler, 2013; Muñoz and Viglianti, 2015). In doing so three main views were chosen: (i) a digital reconstruction of the manuscript (*topographical view*), (ii) a diplomatic transcription (*document-focused view*), and (iii) a normalized, historical-critical version of the text (*text-focused view*).

- **Topographical view**: The topographical view includes the digital reproduction of the original as a two-dimensional, detailed representation of the surface. In other words, it means that the earlier text of the manuscript, significantly enlarged, has been digitally replenished with stylus on the touchscreen (see image 1 and 2). To my knowledge this method has not been used in the reconstruction of the palimpsests before and therefore represents a novel way of displaying original texts. Its uniqueness lies in the combination of human expertise and the current technical possibilities.
- **Document-focused view**: In this view, the text is reproduced as true to the original as possible. The design of the original is visualized, especially the layout of the text and the navigation elements in it (see images 2 and 3). Wherever possible, the original orthography is reproduced. Nevertheless it is also possible to switch to the normalized orthography within the same view. The reader can therefore choose between modern and medieval punctuation, between the spelling containing original abbreviations, or expanded text, etc.
- **Text-focused view** (see image 4): This view has the layout of a modern edition, the orthography is largely normalized. An option is provided to highlight different types of content, such as quotations, places, personal names, peoples, etc.

Technically the concept of the edition is implemented as follows. The topographical view is produced as images. The transcription of the original text for the document- and text-focused views is made according to the TEI/XML standards. Phenomena encoded with tags are divided into larger *blocks*. The main blocks are:

- *Physical condition of the manuscript and physical structure of the text*: physical damage of the manuscript, legibility of the text, page and line division at the locations where they are not connected with the logical structure of the text (see below);
- *Logical structure of the text, as well as all the elements of the layout, which support navigation in the text*: units such as volumes, chapters, excerpts; elements of the design that refer to this classification (e.g. larger spaces in text intentionally left blank); pictograms and marginal notes;
- *Orthography of the manuscript*: original punctuation, accentuation, abbreviations and ligatures;
- *Normalized orthography*: modern punctuation, word boundaries (missing in the manuscript), uppercase and lowercase letters according to modern standards;
- *Contents of the text*: quotes, names, places, peoples, and other items.

The web presentation is created on the basis of XSLT. The transformation of separate blocks of tags is modeled for each view individually. The final presentation is styled with Cascading Style Sheets (CSS). It is planned to publish the edition online by mid-2016.

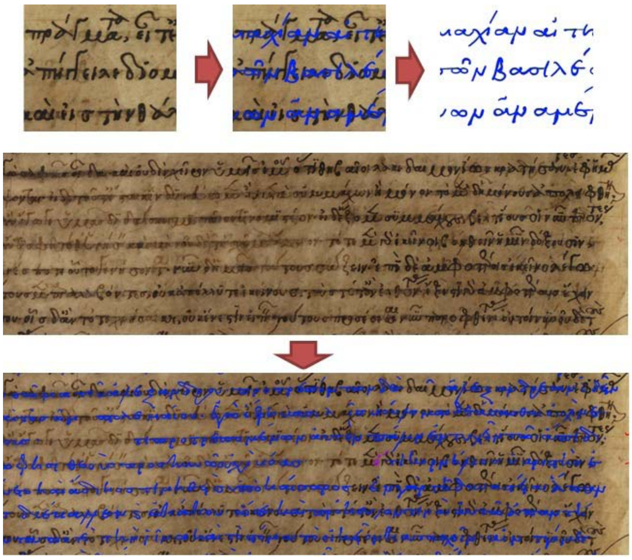


Image 1: The process of the graphical reconstruction of the palimpsest

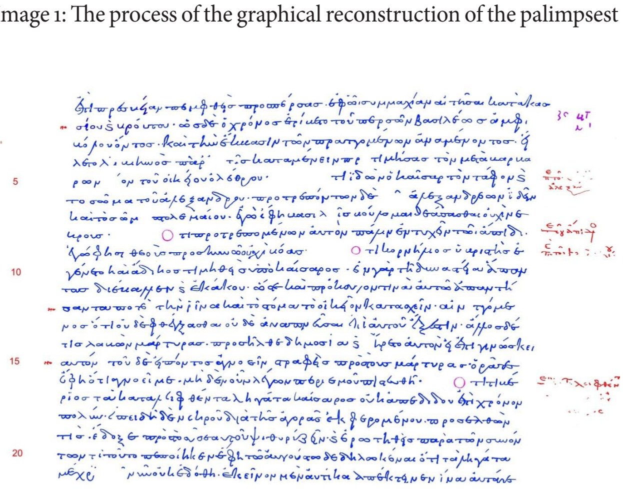
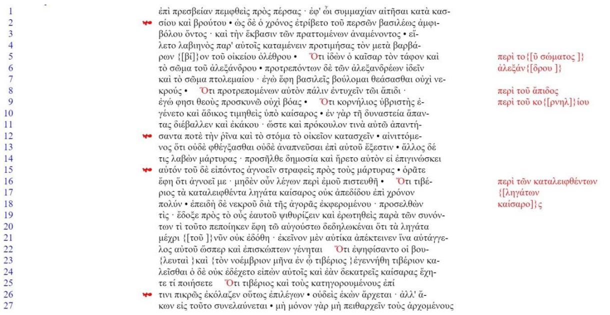
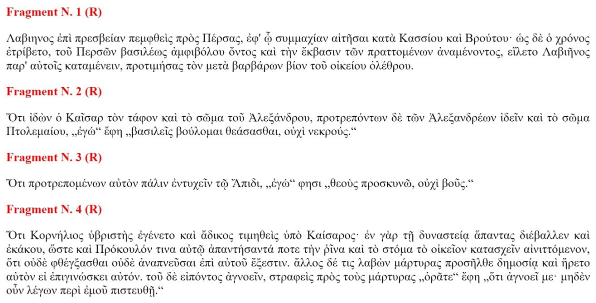


Image 2: A screenshot of the topographical view



ἐπὶ πρεσβείαν πεμφθεὶς πρὸς πέρσας · ἐφ' ὧι συμμαχίαν αἰτῆσαι κατὰ κασ-
σίου καὶ βρούτου • ὡς δὲ ὁ χρόνος ἐτρίβετο τοῦ περσῶν βασιλέως ἀμφι-
βόλου ὄντος · καὶ τὴν ἔκβασιν τῶν πραττομένων ἀναμένοντος • εἵ-
λετο λαβινὸς παρ' αὐτοῖς καταμένειν προτιμήσας τὸν μετὰ βαρβά-
ρων {[βί]}ον τοῦ οἰκείου ὀλέθρου • Ὅτι ἰδὼν ὁ καῖσαρ τὸν τάφον καὶ περὶ το{[ῦ σώματος]}
τὸ σῶμα τοῦ ἀλεξάνδρου • προτρεπόντων δὲ τῶν ἀλεξανδρέων ἰδεῖν ἀλεξάν{[δρου]}
καὶ τὸ σῶμα πτολεμαίου · ἐγὼ ἔφη βασιλεῖς βούλομαι θεάσασθαι οὐχὶ νε-
κρούς • Ὅτι προτρεπομένων αὐτὸν πάλιν ἐντυχεῖν τῶι ἄπιδι · περὶ τοῦ ἄπιδος
ἐγώ φησι θεοὺς προσκυνῶ οὐχὶ βόας • Ὅτι κορνήλιος ὑβριστὴς ἐ- περὶ τοῦ κο{[ρνηλ]}ίου
γένετο καὶ ἄδικος τιμηθεὶς ὑπὸ καίσαρος • ἐν γὰρ τῆ δυναστεία ἅπαν-
τας διέβαλλεν καὶ ἐκάκου · ὥστε καὶ πρόκουλον τινὰ αὐτῶ ἀπαντή-
σαντα ποτὲ τὴν ῥῖνα καὶ τὸ στόμα τὸ οἰκεῖον κατασχεῖν • αἰνιττόμε-
νος ὅτι οὐδὲ φθέγξασθαι οὐδὲ ἀναπνεῦσαι ἐπὶ αὐτοῦ ἔξεστιν • ἄλλος δέ
τις λαβὼν μάρτυρας · προσῆλθε δημοσία καὶ ἤρετο αὐτὸν εἰ ἐπιγινώσκει
αὐτόν τοῦ δὲ εἰπόντος ἀγνοεῖν στραφεὶς πρὸς τοὺς μάρτυρας • ὁρᾶτε
ἔφη ὅτι ἀγνοεῖ με · μηδὲν οὖν λέγων περὶ ἐμοῦ πιστευθῆ • Ὅτι τιβέ- περὶ τῶν καταλειφθέντων
ριος τὰ καταλειφθέντα ληγάτα καίσαρος οὐκ ἀπεδίδου ἐπὶ χρόνον {[ληγάτων
πολύν • ἐπειδὴ δὲ νεκροῦ διὰ τῆς ἀγορᾶς ἐκφερομένου · προσελθὼν καίσαρο]}ς
τὶς · ἔδοξε πρὸς τὸ οὖς ἑαυτοῦ ψιθυρίζειν καὶ ἐρωτηθεὶς παρὰ τῶν συνόν-
των τὶ τοῦτο πεποίηκεν ἔφη τῶ αὐγούστω δεδηλωκέναι ὅτι τὰ ληγάτα
μέχρι {[τοῦ]}νῦν οὐκ ἐδόθη · ἐκεῖνον μὲν αὐτίκα ἀπέκτεινεν ἵνα αὐτάγγε-
λος αὐτοῦ ὥσπερ καὶ ἐπισκώπτων γένηται Ὅτι ἐψηφίσαντο οἱ βου-
{λευταὶ }καὶ {τὸν νοέμβριον μῆνα ἐν ᾧ τιβέριος }ἐγεννήθη τιβέριον κα-
λεῖσθαι ὁ δὲ οὐκ ἐδέχετο εἰπὼν αὐτοῖς καὶ ἐὰν δεκατρεῖς καίσαρας ἔχη-
τε τί ποιήσετε Ὅτι τιβέριος καὶ τοὺς κατηγορουμένους ἐπί
τινι πικρῶς ἐκόλαζεν οὕτως ἐπιλέγων • οὐδεὶς ἑκὼν ἄρχεται · ἀλλ' ἄ-
κων εἰς τοῦτο συνελαύνεται • μὴ μόνον γὰρ μὴ πειθαρχεῖν τοὺς ἀρχομένους

Image 3: A screenshot of the document-focused view



Fragment N. 1 (R)

Λαβιηνος ἐπὶ πρεσβείαν πεμφθεὶς πρὸς Πέρσας, ἐφ' ᾧ συμμαχίαν αἰτῆσαι κατὰ Κασσίου καὶ Βρούτου· ὡς δὲ ὁ χρόνος ἐτρίβετο, τοῦ Περσῶν βασιλέως ἀμφιβόλου ὄντος καὶ τὴν ἔκβασιν τῶν πραττομένων ἀναμένοντος, εἵλετο Λαβιῆνος παρ' αὐτοῖς καταμένειν, προτιμήσας τὸν μετὰ βαρβάρων βίον τοῦ οἰκείου ὀλέθρου.

Fragment N. 2 (R)

Ὅτι ἰδὼν ὁ Καῖσαρ τὸν τάφον καὶ τὸ σῶμα τοῦ Ἀλεξάνδρου, προτρεπόντων δὲ τῶν Ἀλεξανδρέων ἰδεῖν καὶ τὸ σῶμα Πτολεμαίου, „ἐγώ“ ἔφη „βασιλεῖς βούλομαι θεάσασθαι, οὐχὶ νεκρούς.“

Fragment N. 3 (R)

Ὅτι προτρεπομένων αὐτὸν πάλιν ἐντυχεῖν τῷ Ἄπιδι, „ἐγώ“ φησι „θεοὺς προσκυνῶ, οὐχὶ βοῦς.“

Fragment N. 4 (R)

Ὅτι Κορνήλιος ὑβριστὴς ἐγένετο καὶ ἄδικος τιμηθεὶς ὑπὸ Καίσαρος· ἐν γὰρ τῇ δυναστείᾳ ἅπαντας διέβαλλεν καὶ ἐκάκου, ὥστε καὶ Πρόκουλόν τινα αὐτῷ ἀπαντήσαντά ποτε τὴν ῥῖνα καὶ τὸ στόμα τὸ οἰκεῖον κατασχεῖν αἰνιττόμενον, ὅτι οὐδὲ φθέγξασθαι οὐδὲ ἀναπνεῦσαι ἐπὶ αὐτοῦ ἔξεστιν. ἄλλος δέ τις λαβὼν μάρτυρας προσῆλθε δημοσίᾳ καὶ ἤρετο αὐτὸν εἰ ἐπιγινώσκει αὐτόν. τοῦ δὲ εἰπόντος ἀγνοεῖν, στραφεὶς πρὸς τοὺς μάρτυρας „ὁρᾶτε“ ἔφη „ὅτι ἀγνοεῖ με· μηδὲν οὖν λέγων περὶ ἐμοῦ πιστευθῇ.“

Image 4: A screenshot of the text-focused view

Bibliography

Boissevain, U. Ph. (1906). Excerpta historica iussu Constantini Porphyrogeniti confecta. Vol. **4**, *Excerpta de sententiis*. Berlin.

Büttner-Wobst, Th. (1906). Excerpta historica iussu Constantini Porphyrogeniti confecta. Vol. **2**(1), *Excerpta de virtutibus et vitiis*. Berlin.

De Boor, C. G. (1903). Excerpta historica iussu Constantini Porphyrogeniti confecta. Vol.**2**, *Excerpta de legationibus*. Berlin.

De Boor, C. G. (1905). Excerpta historica iussu Constantini Porphyrogeniti confecta. Vol. **3**, *Excerpta de insidiis*. Berlin.

Muñoz, T. and Viglianti, R. (2015). Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive, *Journal of the Text Encoding Initiative* [Online], Issue 8 - PREVIEW | 2014-2015. URL: http://jtei.revues.org/1270.

Pierazzo, E. and Stokes, P. A. (2011). Putting the Text back into Context: A Codicological Approach to Manuscript Transcription, *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age* 2. Schriften des Instituts für Dokumentologie und Editorik, 3. Books on Demand (BoD), Norderstedt, pp. 397-429.

Rehbein, M. and Gabler, H. W. (2013). On Reading Environments for Genetic Editions, *Scholarly and Research Communication* **4**(3). http://src-online.ca/index.php/src/article/view/123 (accessed 10 October 2015).

Roos, A. G. (1910). Excerpta historica iussu Constantini Porphyrogeniti confecta. Vol. **2**(2), *Excerpta de virtutibus et vitiis*. Berlin.

Mapping Imagined and Experienced Places: An Exploration of the Geography of Willa Cather's Writing

Emily J. Rau
emilyjanerau@gmail.com
University of Nebraska-Lincoln, United States of America

Gabi Kirilloff
gkirilloff@gmail.com
University of Nebraska-Lincoln, United States of America

Introduction

The American novelist, Willa Cather, was an extensive traveler whose experience of and attachment to places greatly influenced her writing. As an author whose novels are traditionally regarded as depicting fictional versions of real places, Cather's novels are particularly suited for an examination of the artistic relationship between real and imagined places. The effects that geography had on Cather's work are demonstrated not only by her fiction, but also by her recently published letters, many of which include vivid descriptions of geographic places. This project explores the relationships between real places and those depicted in Cather's writing by creating a series of digital, interactive maps. Using both text analysis and digitally cartographic tools, we examine the differences between Cather's first-hand experience of place and her mental conception of place.

Methodologies

Our project employs two distinct approaches in order to gain a better understanding of the representations of place in Cather's works; the first approach is concerned with Cather's references to real locations and the second explores fictional places. We created digital maps using ArcGIS, a geographic information system designed to capture and analyze geographical data. The first series of maps we created examine place on a global scale. They provide readers and scholars with a visual representation of the places Cather lived in and traveled to, the places she wrote about in 200 of her letters, and the places she references in her novel *My Ántonia*. In order to create these maps, we first had to extract spatial data from the correspondence and *My Ántonia*, which we accomplished by using preexisting information encoded in the Willa Cather Archive's XML and by using the Stanford Named Entity Recognizer.

The second series of maps examine space on a more intimate scale. These maps compare Cather's fictional depictions of the town of Red Cloud Nebraska with the geographic layout of the town at the time Cather experienced it in the late 19th century. In this project, we used ArcGIS to create a map of Red Cloud, NE, circa 1890, based on data provided by the Webster County Assessor's Office and historic information gathered from the Nebraska State Historical Society. The map is interactive, so that each location includes historical information and a photo, when available. From this historic and geographic data, we constructed interpretive maps of two fictional versions of Red Cloud found in Cather's fiction: Black Hawk of *My Ántonia* (1918) and Sweet Water of *A Lost Lady* (1923).

Results

The maps we created helped us to expose Cather's engagement with real space, highlighting moments when she departs from mappable, physical space into the realm of imagination. Further, by uncovering these intricacies and depicting them in maps, readers can perform interpretive work concerning Cather's production of fictional space. The world map of *My Ántonia* indicates that cities within Nebraska are mentioned frequently, as are Eastern European countries. This result makes sense since the novel centers on Eastern European immigrants living in Nebraska. However, more granular Eastern European locations, such as specific cities, do not appear in the novel. Surprisingly, the majority of European cities mentioned are in Italy. This suggests a correlation between Cather's personal experience and the places she writes about, as she visited Italy, but never visited Eastern Europe. Similarly, the local map of *My Ántonia* highlights Cather's reliance on personal experience, as it correlates closely to the historic map of Red Cloud. Scholars have often emphasized that Cather based Jim Burden's home on her grandparents' house, in the northwest part of town. However, the map of the novel reveals that Jim's home is beside the Harlings' house in the southwest part of town, which is the location of Cather's childhood home. In the novel, Cather transplants the internal details of her grandparents' home into the location of her childhood home.

Future Work

The eventual goal of this project is to incorporate all of Cather's letters and novels into our maps and to create additional maps of Cather's other fictional depictions of real places. Expanding the texts used would provide a more nuanced picture of the places Cather writes about. In addition, including more texts will allow users to better query the data, asking questions such as, which of Cather's novels include the most references to places she experienced first-hand? Do Cather's fictional depictions of Pittsburgh and New York align as closely with reality as her re-imaginations of historic Red Cloud? Creating additional maps will allow for an exploration of the ways in which Cather's frame of spatial reference shifts among these works.

CTRaCE: Canonical Text Reader and Citation Exporter

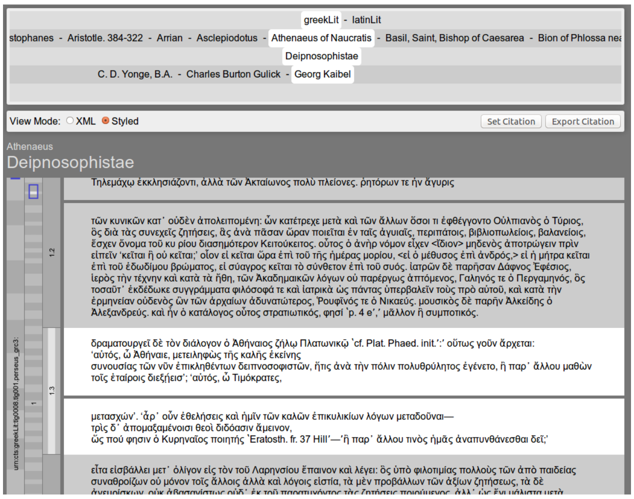
Martin Reckziegel
reckziegel@informatik.uni-leipzig.de
Leipzig University, Germany

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

Gerik Scheuermann
scheuermann@informatik.uni-leipzig.de
Leipzig University, Germany

Motivation

"Citation is the core of scholarship" Neel Smith and Chris Blackwell state in their paper on the CTS/CITE Architecture (Smith et al., 2012). A citation like "Athenaeus, *Deipnosophists*, edition of Kaibel (1887), book 1, chapter 3" can refer to only one string of Greek words and it is valid for every printed copy of Kaibel's edition. Such a citation identifies an object of study uniquely, and it is "independent of any implementing technology". But such a citation scheme is hardly machine readable, and the digital processing of citations becomes more and more important in the humanities. To support this task, the CITE Architecture 1 was designed. Adopting the international standard of URNs 2, unique, complete and machine readable scholarly citations, so called CTS URNs can be generated.


greekLit - latinLit
stophanes - Aristotle. 384-322 - Arrian - Asclepiodotus - Athenaeus of Naucratis - Basil, Saint, Bishop of Caesarea - Bion of Phlossa nea
Deipnosophistae
C. D. Yonge, B.A. - Charles Burton Gulick - Georg Kaibel
View Mode: ○ XML ● Styled
Set Citation
Export Citation
Athenaeus
Deipnosophistae
urn:cts:greekLit:tlg0008.tlg001.perseus_grc3:
1
1.2
1.3

Τηλεμάχῳ ἐκκλησιάζοντι, ἀλλὰ τῶν Ἀκταίωνος πολὺ πλείονες. ῥητόρων τε ἠν ἄγυρις

τῶν κυνικῶν κατ᾽ οὐδὲν ἀπολειπομένη: ὧν κατέτρεχε μετὰ καὶ τῶν ἄλλων ὅσοι τι ἐφθέγγοντο Οὐλπιανὸς ὁ Τύριος, ὃς διὰ τὰς συνεχεῖς ζητήσεις, ἃς ἀνὰ πᾶσαν ὥραν ποιεῖται ἐν ταῖς ἀγυιαῖς, περιπάτοις, βιβλιοπωλείοις, βαλανείοις, ἔσχεν ὄνομα τοῦ κυ ρίου διασημότερον Κειτούκειτος. οὗτος ὁ ἀνὴρ νόμον εἶχεν <ἴδιον> μηδενὸς ἀποτρώγειν πρὶν εἰπεῖν 'κεῖται ἢ οὐ κεῖται;' οἷον εἰ κεῖται ὥρα ἐπὶ τοῦ τῆς ἡμέρας μορίου, <εἰ ὁ μέθυσος ἐπὶ ἀνδρός,> εἰ ἡ μήτρα κεῖται ἐπὶ τοῦ ἐδωδίμου βρώματος, εἰ σύαγρος κεῖται τὸ σύνθετον ἐπὶ τοῦ συός. ἰατρῶν δὲ παρῆσαν Δάφνος Ἐφέσιος, ἱερὸς τὴν τέχνην καὶ κατὰ τὰ ἤθη, τῶν Ἀκαδημαικῶν λόγων οὐ παρέργως ἁπτόμενος, Γαληνός τε ὁ Περγαμηνός, ὃς τοσαῦτ᾽ ἐκδέδωκε συγγράμματα φιλόσοφά τε καὶ ἰατρικὰ ὡς πάντας ὑπερβαλεῖν τοὺς πρὸ αὐτοῦ, καὶ κατὰ τὴν ἑρμηνείαν οὐδενὸς ὢν τῶν ἀρχαίων ἀδυνατώτερος, Ῥουφῖνός τε ὁ Νικαεύς. μουσικὸς δὲ παρῆν Ἀλκείδης ὁ Ἀλεξανδρεύς. καὶ ἦν ὁ κατάλογος οὗτος στρατιωτικός, φησί ʽp. 4 e′,′ μᾶλλον ἢ συμποτικός.

δραματουργεῖ δὲ τὸν διάλογον ὁ Ἀθήναιος ζήλῳ Πλατωνικῷ ʽcf. Plat. Phaed. init.′:′ οὕτως γοῦν ἄρχεται:
'αὐτός, ὦ Ἀθήναιε, μετειληφὼς τῆς καλῆς ἐκείνης
συνουσίας τῶν νῦν ἐπικληθέντων δειπνοσοφιστῶν, ἥτις ἀνὰ τὴν πόλιν πολυθρύλητος ἐγένετο, ἢ παρ᾽ ἄλλου μαθὼν τοῖς ἑταίροις διεξῄεισ'; 'αὐτός, ὦ Τιμόκρατες,

μετασχών'. 'ἆρ᾽ οὖν ἐθελήσεις καὶ ἡμῖν τῶν καλῶν ἐπικυλικίων λόγων μεταδοῦναι—
τρὶς δ᾽ ἀπομαξαμένοισι θεοὶ διδόασιν ἄμεινον,
ὥς πού φησιν ὁ Κυρηναῖος ποιητής ʽEratosth. fr. 37 Hill′—′ἢ παρ᾽ ἄλλου τινὸς ἡμᾶς ἀναπυνθάνεσθαι δεῖ;'

εἶτα εἰσβάλλει μετ᾽ ὀλίγον εἰς τὸν τοῦ Λαρηνσίου ἔπαινον καὶ λέγει: ὃς ὑπὸ φιλοτιμίας πολλοὺς τῶν ἀπὸ παιδείας συναθροίζων οὐ μόνον τοῖς ἄλλοις ἀλλὰ καὶ λόγοις εἱστία, τὰ μὲν προβάλλων τῶν ἀξίων ζητήσεως, τὰ δὲ ἀνευρίσκων, οὐκ ἀβασανίστως οὐδ᾽ ἐκ τοῦ παρατυχόντος τὰς ζητήσεις ποιούμενος, ἀλλ᾽ ὡς ἔνι μάλιστα μετὰ

Figure 1: Screenshot of *CTRaCE* showing the CTS URN greekLit:tlg0008.tlg001.perseus_grc3:1.3:

Although the CITE Architecture is a substantial basis, it does not provide an interface to intuitively make this capability accessible to humanities scholars. This paper aims to fill this gap. We designed the web-based application *CTRaCE* that first of all allows the browsing and reading of texts provided in the corresponding canonical format. With basic interaction functionality, *CTRaCE* supports easy creation of a digital citation of an arbitrary text passage as a CTS URN. As it can also be used to resolve existent CTS URNs, *CTRaCE* fulfills the major requirements for the processing of manually generated digital citations necessary for digital scholarship.

Technical Basis

CTRaCE is based on Canonical Text Services (CTS), which – part of the CITE architecture – addresses two fundamental needs in digital scholarship: (1) the citation of textual units, and (2) their retrieval. To accomplish the transition between classical and digital citations, CTS use Uniform Resource Names (URNs). The top-level structure of a CTS URN 3 has the following format:

cts:urn:namespace:work:passage

As URNs are organized hierarchically from left to right, each component increases the precision of the corresponding reference. The *namespace* component is the top level division of the system. As we use the Perseus Digital Library 4 as demonstration corpus, one of the namespaces is "greekLit", which identifies works in ancient Greek that are preserved through manuscript tradition (Smith et al., 2012). The *work* component specifies the document the URN refers to. It is divided into *textgroup.work.version.exemplar* from which the last two parts are optional. For example, the URN

cts:urn:greekLit:tlg0008.tlg001.perseus_grc3:

refers to Kaibel's edition of "Deipnosophists" (tlg001), located in the textgroup that contains works by Athenaeus (tlg0008). Generating a URN representation for each text, the CTS architecture organizes the whole corpus in the form of a tree with inner nodes representing the hierarchical structure among the texts, which are the leaf nodes of the tree.

Each individual text further has a canonical citation scheme assigned. As the hierarchy of texts varies depending on genre – a play might be arranged by acts and scenes while a poem uses verses –, this scheme is flexibly designed. Additionally, the URN can be further refined with the *passage* component that specifies a continuous part of the text – even on character granularity if necessary.

cts:urn:greekLit:tlg0008.tlg001.perseus_grc3:1.2@καὶ[2]-1.3@δρ

refers to the "Deipnosophists" part that ranges from the second occurrence of the string "καὶ" in the second chapter of the first book until the first occurrence of "δρ" in the third chapter.

Figure 2 displays an overview of our CTS based architecture. Though the scalable CTS implementation (Tiepmar et al., 2014) we use as server infrastructure supports any text in Unicode, we built our application with the idea to operate with texts in TEI 5 format that often include styling markup, which can be used for visualization purposes. *CTRaCE* is a client side web application communicating with the CTS server using the CTS protocol 6. In the following section, we explain the design and the functionality of *CTRaCE* in detail.

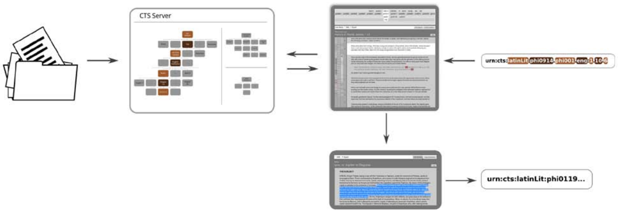

CTS Server
urn:cts:latinLit:phi0914.phi001.eng:1.10.6
urn:cts:latinLit:phi0119...

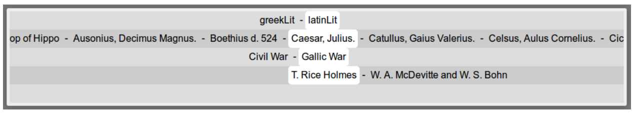
Figure 2: Architecture overview

User Interface

The CTS architecture is a robust basis for digital citations, but due to its technical nature, tools need to be built upon the architecture to leverage its functionality in an intuitive way to humanities scholars. We present *CTRaCE* for that purpose, an example screenshot is shown in Figure 1. It focuses on supporting the following three tasks:

- browsing and reading texts in the CTS system
- creating and exporting citations using CTS URNs
- resolving CTS URNs by visualizing the cited text and its surrounding context

CTRaCE is designed to display the content of a single URN, always shown as parameter in the address bar of the internet browser. Everything related to the citation this URN points to is drawn with a white background in three view components, which we explain below.



greekLit - latinLit

op of Hippo - Ausonius, Decimus Magnus. - Boethius d. 524 - Caesar, Julius. - Catullus, Gaius Valerius. - Celsus, Aulus Cornelius. - Cic

Civil War - Gallic War

T. Rice Holmes - W. A. McDevitte and W. S. Bohn

Figure 3: Navigation component

The **navigation component** shown in Figure 3 is placed in the header section of the user interface and can be used to browse through all available documents. It is derived from the ontology of the CTS system described above. Each row represents one level of the URN parts that identify a text. If appropriate metadata such as authors or titles are available, we display these labels instead of URN identifiers as fallback. Each text in this tree can be selected via mouse click for close reading.

CTRaCE displays texts as one continuous scrollable block, which for optimal performance is loaded in chunks. To the left of the text, we implemented a **distant view component** showing the citation structure in relation to the visible portion of the text. We therefore adopt a visualization method called Icicle Plot (Kruskal et al, 1983) that shows hierarchical relations in a compact way. Figure 4 shows a screenshot of this view. The underlying text uses a three level hierarchy to create a citation: book, chapter and section. Each gray-bordered rectangle in the view represents a node within the citation structure. The leftmost rectangle labeled with the URN of the document represents the root. Inside each node's rectangle, we provide information about their direct children regarding the citation tree. Since the document contains eight books, the same number of alternating shaded rectangles is drawn inside the root node. We provide quick navigation through the document using mouse clicks on these inner rectangles. Depending on the document position, the corresponding citation nodes are drawn in the remaining columns with the leaf nodes aligned to the text. As the text shown in Figure 4 is only part of the first chapter, there is only one node rectangle in the second column. Additionally, the white rectangles depict which part of a node is defined by the current URN, and the blue rectangles represent the visible portion of a node.

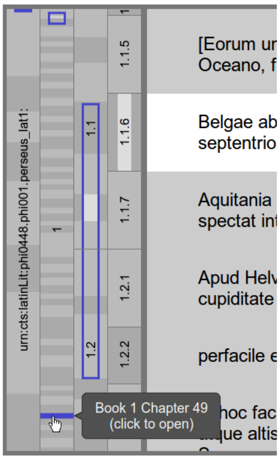
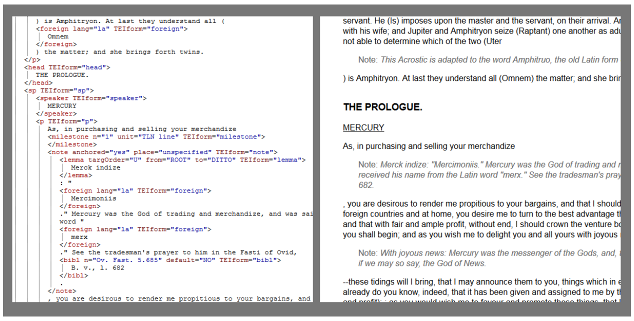


Figure 4: Distant view component

The text of the document is shown in the **text reader component** in main area of the screen. As the text chunks are loaded from the server, they are transformed into HTML code. So far we implemented two rendering methods as shown in Figure 5. The XML view mode displays the text including its markup in a syntax highlighted, indented fashion. The Styled view mode only displays the text using the markup information to style it appropriately. For example, <head> tags will be rendered bold, <note> tags indented, etc. The user can switch between both modes while viewing the document.

When generating a new digital citation with our application, basically the current URN is changed. There are two possibilities to do so. First, the scholar can click on one of the labels of the nodes in the distant view component. This selects the corresponding entire citation node.

Second, the user can mark the desired text passage in the text reader component with the mouse, and then click the "Set Citation" button. The updated URN then represents the selected text passage. Finally, a popup dialog helps to export the URN or the text it is referring to. This includes the possibility to generate a link pointing to our web application showing the generated citation.



```
    ) is Amphitryon. At last they understand all (
    <foreign lang="la" TEIform="foreign">
      Omnem
    </foreign>
    ) the matter; and she brings forth twins.
</p>
<head TEIform="head">
  THE PROLOGUE.
</head>
<sp TEIform="sp">
  <speaker TEIform="speaker">
    MERCURY
  </speaker>
  <p TEIform="p">
    As, in purchasing and selling your merchandize
    <milestone n="1" unit="TLN line" TEIform="milestone">
    </milestone>
    <note anchored="yes" place="unspecified" TEIform="note">
      <lemma targOrder="U" from="ROOT" to="DITTO" TEIform="lemma">
        Merck indize
      </lemma>
      : "
      <foreign lang="la" TEIform="foreign">
        Mercimoniis
      </foreign>
      ." Mercury was the God of trading and merchandize, and was sai
      word "
      <foreign lang="la" TEIform="foreign">
        merx
      </foreign>
      ." See the tradesman's prayer to him in the Fasti of Ovid,
      <bibl n="Ov. Fast. 5.685" default="NO" TEIform="bibl">
        B. v., l. 682
      </bibl>
      .
    </note>
    , you are desirous to render me propitious to your bargains, and
```

servant. He (Is) imposes upon the master and the servant, on their arrival. Ar
with his wife; and Jupiter and Amphitryon seize (Raptant) one another as adu
not able to determine which of the two (Uter

Note: *This Acrostic is adapted to the word Amphitruo, the old Latin form*

) is Amphitryon. At last they understand all (Omnem) the matter; and she brir

THE PROLOGUE.

MERCURY

As, in purchasing and selling your merchandize

Note: *Merck indize: "Mercimoniis." Mercury was the God of trading and r*
received his name from the Latin word "merx." See the tradesman's pray
682.

, you are desirous to render me propitious to your bargains, and that I should
foreign countries and at home, you desire me to turn to the best advantage th
and that with fair and ample profit, without end, I should crown the venture bo
you shall begin; and as you wish me to delight you and all yours with joyous

Note: *With joyous news: Mercury was the messenger of the Gods, and,*
if we may so say, the God of News.

--these tidings will I bring, that I may announce them to you, things which in e
already do you know, indeed, that it has been given and assigned to me by th

Figure 5: XML (left) and Styled (right) rendering methods

Conclusion

The proposed *CTRaCE* interface is based on Canonical Text Services in order to provide an online service to generate citations of digital text editions. Such a front-end allows humanities scholars not only to close read texts, but also to identify and retrieve text passages of interest. The interface offers the possibility to generate and export citations of textual works down to the character level. Moreover, *CTRaCE* is a service that is capable of resolving existent or with *CTRaCE* generated citations by opening the referenced text passage.

As of now, *CTRaCE* is a very helpful tool for humanities scholars who want to produce machine readable citations. Future steps in the development of *CTRaCE* include visualizations of different kinds of annotations on a CTS passage produced by different scholars to facilitate the collaborative work on a shared text of interest.

Acknowledgments

Parts of the work presented here is the result of the project "Die Bibliothek der Milliarden Wörter" funded by the European Social Fund. We additionally want to thank Monica Berti for the valuable input and discussion about use cases for our application.

Notes

[1] http://www.homermultitext.org/hmt-doc/cite/
[2] https://www.ietf.org/rfc/rfc2141.txt
[3] https://github.com/cite-architecture/ctsurn_spec/blob/master/md/specification.md
[4] https://github.com/PerseusDL/canonical
[5] http://www.tei-c.org/index.xml
[6] https://github.com/cite-architecture/cts_spec/blob/master/md/specification.md

Bibliography

Kruskal, J. B. and Landwehr, J. M. (1983). Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, **37**(2): 162-68.

Smith, D. N. and Blackwell, C. (2012). Four URLs, Limitless Apps: Separation of concerns in the Homer Multitext Architecture. In Muellner, L. (ed.), *Donum natalicium digitaliter confectum Gregorio Nagy septuagenario a discipulis collegis familiaribus oblatum*, Washington, DC: The Center for Hellenic Studies of Harvard University.

Tiepmar, J., Teichmann, C., Heyer, G., Berti, M., and Crane, G. (2014). A New Implementation for Canonical Text Services. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Association for Computational Linguistics.

A Tool for NLP-Preprocessing in Literary Text Analysis

Nils Reimers
reimers@ukp.informatik.tu-darmstadt.de
TU Darmstadt, Ubiquitous Knowledge Processing Lab

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Wuerzburg, Germany

Stefan Pernes
stefan.pernes@uni-wuerzburg.de
University of Wuerzburg, Germany

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Wuerzburg, Germany

Isabella Reger
isabella.reger@uni-wuerzburg.de
University of Wuerzburg, Germany

Thorsten Vitt
thorsten.vitt@uni-wuerzburg.de
University of Wuerzburg, Germany

The possibilities for widening the spectrum of research questions by adopting new computational methodology seem to be almost unlimited for literary scholars with considerable programming skills. Researchers with little or no such skills, however, have to rely on user-friendly tools. Simple word counts are still among the most common, and admittedly often very useful features used in computational text analysis. Usually, linguistic annotations are needed for using more complex features in the analysis of style or content of a literary text. For example, a researcher might want to investigate style in terms of syntactic preferences by applying stylometric analysis on part-of-speech tag n-grams, to run topic modelling on specific word types only or to characterize the way an author describes figures by extracting all the adjectives that refer to a named entity. All these examples require of the scholar to first extract linguistic information from the text and use that information to define complex features.

Computer linguists have developed several tools for the various tasks of natural language processing (NLP) that can automatically analyze a digital text and annotate it with such information. In the present spectrum of solutions for NLP tasks, there is a gap between tools for rather simple tasks and full programming frameworks which require significant programming skills. The one end of the spectrum is represented by WebLicht,[1] a web service that allows users to upload and process single files very comfortably. On the other end are GATE,[2] NLTK[3], BookNLP[4] and the Darmstadt Knowledge Processing Repository (DKPro).[5]

DKPro provides a programming framework in which many such NLP tools can be combined into an analysis pipeline. The pipelining approach is especially useful, often even necessary, when one NLP tool needs the annotations provided by another NLP tool in advance for extracting more complex linguistic features. DKPro thus provides access to tools like sentence splitters, tokenizers, part-of-speech taggers, named-entity recognizers, lemmatizers, morphological analyzers and parsers in many languages.[6]

While making NLP significantly easier by integrating many NLP tools into a single framework, the use of DKPro still requires a substantial knowledge of technologies like UIMA, Java and Maven. To further lower the skill threshold for literary scholars to use complex NLP output in computational text analysis, DARIAH-DE (the German branch of the European project Digital Research Infrastructure for the Arts and Humanities, funded by the German Federal Ministry of Education and Research) developed the DARIAH-DKPro-Wrapper (DDW).[7] The DDW bundles a pipeline with a set of commonly used NLP components into a java program to be executed with a single command. As DKPro in general, the wrapper provides transparent access to a whole set of different NLP tools which are downloaded as needed behind the scenes. Command line options and configuration files allow users a considerable degree of control over the pipeline and its components, giving partial access to DKPro functionality without requiring any programming knowledge. The DDW also solves the problem of different input and output formats of the tools, offering a unified access. Therefore, the DDW positions itself in between the two ends of the aforementioned spectrum: It runs locally, allows for the processing of multiple files and can be configured to a considerable extent to one's own needs. Whereas the user of classical DKPro is a UIMA programmer, the DDW can be used by anybody who can copy a command into the command line. Nonetheless, the DDW in some cases offers more features than other more advanced solutions, as DKPro supports more tools and languages. It also integrates Stanford NLP and supports the highly efficient Treetagger. A list of components available for both DKPro and the DDW can be found of the DKPro project page.

Furthermore, the DDW stores its output in a tab-separated plain text format inspired by CoNLL2009.[8] The format provides information on paragraph id, sentence id, token id, token, lemma, POS, chunk, morphology, named entity, parsing information and more. This format can be comfortably accessed in common scripting languages for further analysis, i.e. it can be directly read as a dataframe object in R or Python Pandas; it can even be opened in a common datasheet editor like Microsoft Excel.

Scripts connecting the output format to popular text analysis tools like the R package *stylo*[9] are currently under development. Dariah also prepared some tutorials explaining how to use the wrapper and showing the use of the output format in research in three use cases.[10]

This poster will present the DDW and its file format as a new and comfortable means of providing linguistic annotations, thus significantly lowering the threshold for using complex NLP-based features in computational literary analysis.

Notes

[1] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

[2] https://gate.ac.uk/

[3] http://www.nltk.org/

[4] https://github.com/dbamman/book-nlp

[5] https://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/

[6] Not every kind of tool is available in all languages; it depends on the native support of the tools, not on the framework provided by DKPro.

[7] https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper

[8] https://ufal.mff.cuni.cz/conll2009-st/task-description.html

[9] Eder, Maciej, Mike Kestemont, and Jan Rybicki. "Stylometry with R: a suite of tools." *Digital Humanities 2013: Conference Abstracts*. 2013. For the software see: https://sites.google.com/site/computationalstylistics/stylo

[10] https://rawgit.com/DARIAH-DE/DARIAH-DKPro-Wrapper/master/doc/tutorial.html

Mapping Languages Performance by Performance

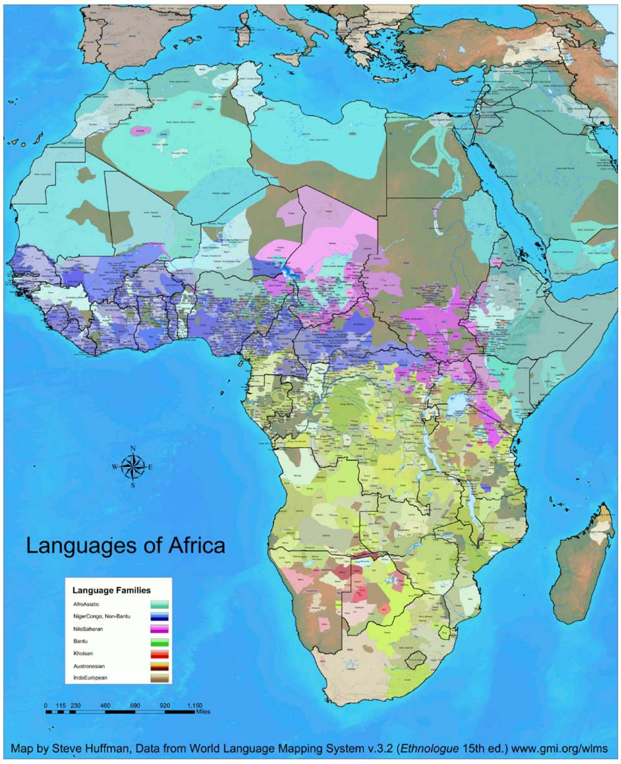
Sandy Ritchie
tr7@soas.ac.uk
SOAS, University of London, United Kingdom

Samantha Goodchild
sg76@soas.ac.uk
SOAS, University of London, United Kingdom

Karolina Grzech
298351@soas.ac.uk
SOAS, University of London, United Kingdom

This presentation introduces Language Landscape (LL) - an online, open access language mapping resource, which adopts a 'performance-based' approach to mapping languages.

Language maps play a crucial role in increasing our understanding of the geographical distribution of linguistic groups (Lameli, 2010). Many language maps represent languages either as polygons or points. Each polygon in Figure 1, or point in Figure 2, represents a language of Africa.



Map by Steve Huffman, Data from World Language Mapping System v.3.2 (*Ethnologue* 15th ed.) www.gmi.org/wlms

Figure 1: polygon-based map of African languages (credit: Steve Huffman/Ethnologue)

While these maps are useful for gaining a general insight into the linguistic makeup of an area, there are issues with the way they represent language communities. Languages are not commonly spoken by a monolingual population who live across a continuous area with discrete boundaries. Language boundaries are fluid, and both individuals and populations are often highly multilingual, for example in some sub-Saharan African settings (Lüpke and Storch, 2013), and also in many large cities (Extra and Barni, 2008). A notional "centre" for a language is also problematic. Speakers of a language are typically dispersed across a large area, meaning that a single point chosen to represent that community can only represent some important cultural centre (Dahl and Veselinova, 2006).

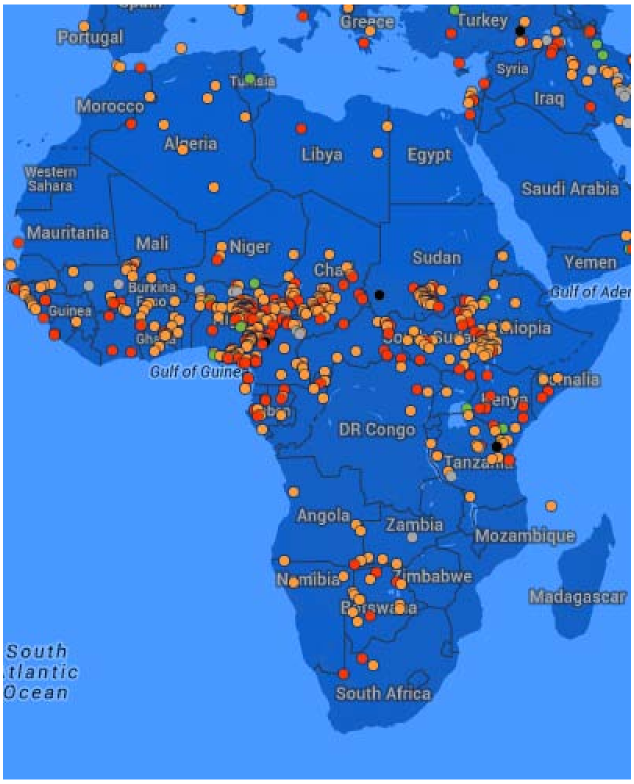


Figure 2: points-based map of African languages (credit: The Endangered Languages Project)

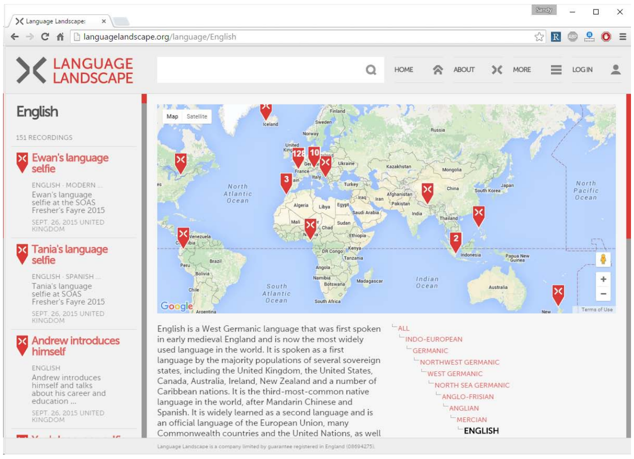
Both polygon- and points-based maps represent languages as monolithic entities which have an existence separate from their speakers, cf. the notion of "competence" (Chomsky, 1965). An alternative way to view languages is as "communities of practice" (e.g. Wenger, 1998; Eckert, 2000) - loose associations of individuals who participate in shared social practices.

One way to represent the spread of communities of practice is to map recordings of community members' performances, because these represent a moment in which speakers participated in a community of practice. Coupled with rich sociolinguistic information about speakers' backgrounds and repertoires, a map of these performances represents not only all the places in which a language is spoken, but also who speaks it, and which other languages those speakers speak.

We have adopted this performance-based approach to language mapping on our website languagelandscape.org.

The website is designed to enable anyone with an internet connection to create an account and add recordings of their languages to a world map. Other users can then access these recordings by browsing or searching the map. In order to encourage as many contributions as possible, we do not have any specific requirements for the type of data submitted. The metadata categories on the website are based on the International Meta Data Initiative (IMDI) standards and are stored in a relational database.

The language maps available on Language Landscape do not represent languages as polygons or single points, but rather as a series of points which indicate the locations where language recordings were made. The maps on LL are also, in this sense, "point-based", since points are used to represent languages, but crucially these points indicate the locations of performances of a language, rather than the language itself. For example, the map below shows recordings of English added to the website, which may be taken as a proxy for (some of the) locations around the world where English is spoken:



English

151 RECORDINGS

Ewan's language selfie
ENGLISH · MODERN ...
Ewan's language selfie at the SOAS Fresher's Fayre 2015
SEPT. 26, 2015 UNITED KINGDOM

Tania's language selfie
ENGLISH · SPANISH ...
Tania's language selfie at SOAS Fresher's Fayre 2015
SEPT. 26, 2015 UNITED KINGDOM

Andrew introduces himself
ENGLISH
Andrew introduces himself and talks about his career and education ...
SEPT. 26, 2015 UNITED KINGDOM

English is a West Germanic language that was first spoken in early medieval England and is now the most widely used language in the world. It is spoken as a first language by the majority populations of several sovereign states, including the United Kingdom, the United States, Canada, Australia, Ireland, New Zealand and a number of Caribbean nations. It is the third-most-common native language in the world, after Mandarin Chinese and Spanish. It is widely learned as a second language and is an official language of the European Union, many Commonwealth countries and the United Nations, as well

ALL
INDO-EUROPEAN
GERMANIC
NORTHWEST GERMANIC
WEST GERMANIC
NORTH SEA GERMANIC
ANGLO-FRISIAN
ANGLIAN
MERCIAN
ENGLISH

Language Landscape is a company limited by guarantee registered in England (08694275).

Figure 3: Language Landscape map of the English language

We believe that given a critical mass of contributors, the website could become a useful tool for a range of potential audiences. Contributions are currently skewed towards places where digital infrastructure is well established. However, technological advances mean that in 5-10 years the network will be available to more people. Resources like our website, which concentrate on promoting audio and video content in minority languages, will play a crucial role in convincing new users that the internet is a useful and valuable tool for communicating remotely.

In order to encourage more contributions, we target students who have expressed an interest in languages through their course choices. We have helped university students in London, Croatia, Germany and Poland to run projects, and have run outreach projects with 12-13 year olds in a school in East London. Promotion of the site on social media has also resulted in members of the public contributing recordings to the site.

We hope that the website will contribute to efforts to increase the number of languages and cultures currently represented on the web. We want to provide a platform where people feel comfortable representing their own language and culture in the way they see fit, rather than feeling compelled to participate in one of the dominant cultural forms which make up so much online content (e.g. Kornai, 2013).

Bibliography

Chomsky, N. A. (1965). *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.

Dahl, Ö. and Veselinova, L. (2006). *Language Map Server. University of Stockholm*. http://www.esri.com/news/arcuser/0206/language_ms1of2.html.

Eckert, P. (2000). *Linguistic Variation as Social Practice*, Malden, MA: Blackwell.

Extra, G. and Barni, M. (2008). Mapping linguistic diversity in multicultural contexts: cross-national and cross-linguistic perspectives. In Barni, M. and Extra, G. (eds), *Mapping Linguistic Diversity in Multicultural Contexts*. Berlin: Walter de Gruyter, pp. 3–42.

Kornai, A. (2013). *Digital language death*. PLoS ONE, **8**(10).

Lameli, A. (2010). Linguistic atlases - traditional and modern. In Auer, P. and Schmidt, J. E. (eds), *Language and Space: An International Handbook of Linguistic Variation*. Berlin, New York: Mouton de Gruyter, pp. 567–92.

Lüpke, F. and Storch, A. (2013). *Repertoires and Choices in African Languages*, Berlin: De Gruyter.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*, Cambridge: Cambridge University Press.

Sussex Humanities Lab – Emotion, Automation and Sonic Socialities

Ben Roberts
b.l.roberts@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Alban Webb
Alban.Webb@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Liam Berriman
L.J.Berriman@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Sharon Webb
Sharon.Webb@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

James Baker
James.Baker@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Beatrice Fazi
B.Fazi@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Andrew Robertson
Andrew.Robertson@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Ben Jackson
Ben.Jackson@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Jack Pay
jp242@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Simon Wibberley
Simon.Wibberley@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

Chris Kiefer
C.Kiefer@sussex.ac.uk
Susex Humanities Lab, University of Sussex, United Kingdom

At the Sussex Humanities Lab we are building collaborations and networks to support and develop new forms of digital humanities. Our vision is to ensure that information scientists, literary theorists, media scholars, designers, practitioners, technologists, philosophers, social scientists and historians collaborate to serve the fundamental roles of humanities research. Sussex Humanities Lab builds on a long standing interdisciplinary culture at Sussex and is developing a programme that considers our digital past and digital futures. This is reflected in our four guiding research strands:

- Digital History and Digital Archives
- Digital Media and Computational Culture
- Digital Lives, Memory and Experience
- Digital Technologies and Digital Performance.

Comprising a team of lecturers, research fellows, PhD students, technicians, and senior faculty, we are interested in developing new research areas across our multi-disciplinary team. Each strand reflects the seed initiatives of the Lab and are a starting point for our research journey. We are dedicated to developing and expanding research into how digital technologies are shaping our culture and society, as well as the way we go about our research. To this end we are developing a number of research projects, grants and network bids. Examples of our current research include:

- **Playing Tag: Identifying Emotion in Oral History Collections:** Although large collections of video and audio files recording oral history testimony have been collected, archived, and aggregated in digital formats, the way we use these files remains largely analogue: We sit and listen and watch them. Research methods from other fields have great potential for the automatic analyses of emotions from postural movements and vocal inflections. By bringing together scholars from the field of oral history and the histories of emotion, social scientists using interview methodologies, sonic and video studies, corpus linguistics, and linguistic anthropology, this research lays the foundations for a new approach to the analyses of oral history collections.
- **Automation Anxiety:** This research explores methods by which the humanities might evaluate contemporary cultural anxiety about automation. From self-driving cars, through high-frequency trading to military drones and organised swarms of shelf-stacking robots, our era is marked by a fascination with a fresh wave of automation technology. This new "rise of the machines" is characterised by the replacement of human decision making with reliance on algorithms, machine learning and other computational techniques whose fitness for purpose cannot be clearly understood by those whose lives they affect (Carr, 2014). The focus of this research is on the cultural anxiety associated with these new technologies and we seek to develop methods, approaches and tools that might be used to analyse and understand it. The question of method is particularly pertinent given that the humanities are themselves being automated. Indeed the debates around digital humanities reproduce much of the anxiety around automation itself. This research addresses the way in which, to paraphrase Ruppert, Law and Savage (2013), automation is increasingly both the *material* of culture and the apparatus for *knowing* that culture.
- **Sonic Ecologies and Socialities:** This research brings together new understandings of how we sensorially engage with sound across digitally mediated bodies and computational environments. Our research is inspired by an eclectic range of perspectives, from performance studies and musicology (Coyne, 2010) to neuroscience, philosophy, sensory ethnography and digital aesthetics (Back 2007; Bidelman and Krishnan, 2009; Fazi and Fuller forthcoming). Common concerns amongst these perspectives are: (1) new modes of sensory perception, the sonic mediation of experience, and the possibility of algorithmically addressing such modes and mediations, (2) the augmentation of bodies and sensory methods for tracing and capturing soundscapes, and (3) new performative and compositional methods for creating and curating soundscapes that incite alternative politics of sensing.

Other activities of the Lab include: visiting fellows scheme, workshops and seminar series.

Bibliography

Back, L. (2003). *The Art of Listening*. London and New York: Bloomsbury.

Bidelman, G. and Krishnan, A. (2009). Neural Correlates of

Consonance, Dissonance, and the Hierarchy of Musical Pitch in the Brainstem. *The Journal of Neuroscience*, **29**(42): 13165-71.

Carr, Nicholas G. (2014). *The Glass Cage: Automation and Us.*

Coyne, R. (2010). *The Tuning of Place: Sociable Spaces and Pervasive Digital Media*. Cambridge, MA.: MIT Press.

Fazi, B. and Fuller, M. (forthcoming), Computational Aesthetics in Paul, C. (Ed) *A Companion to Digital Art*. Wiley and Sons.

Ruppert, E., Law, J. and Savage, M. (2013). Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture and Society*, **30**(4): 22–46.

Textual Communities

Peter Robinson
peter.robinson@usask.ca
University of Saskatchewan, Canada

Barbara Bordalejo
barbara.bordalejo@arts.kuleuven.be
KU Leuven, Belgium

This poster presents the public version of the Textual Communities scholarly editing environment, describing its underlying principles, its innovative structure, and its functionality. It updates the presentation of Textual Communities given as a poster at DH 2013 in Lincoln, Nebraska.

Study of literary works that exist in many different forms is one of the most important and difficult tasks in the humanities. The number of forms a work may have —eighty-four fifteenth-century manuscripts and printed texts of Chaucer's Canterbury Tales, more than eight hundred manuscripts of Dante's Commedia, and five thousand manuscripts of the Greek New Testament — is both testimony to their significance and a challenge to scholars. In order to understand these texts and how they relate, we have to discover as much as we can of how they came to be written and disseminated. Only then can we seek to establish how they might best be read and prepare texts (in the form of scholarly editions) for scholars to use. The work of building archives of primary materials for this work is daunting and prohibitive in the old lone scholar method. Many large editing projects have opted for a team approach, exploring the possibilities of crowd-sourcing for processing large amounts of textual data. The challenge is to coordinate this work in a way that produces high-quality, useful results. The Textual Communities workspace enables every aspect of the scholarly process: defining, transcribing, and linking textual materials for a digital archive or edition; collating the witnesses and adjusting the collation for optimal scholarly use; analyzing the results of the collation to create an understanding of the relations among the manuscripts; and for marshaling and managing a community of participants with an array of community building tools.

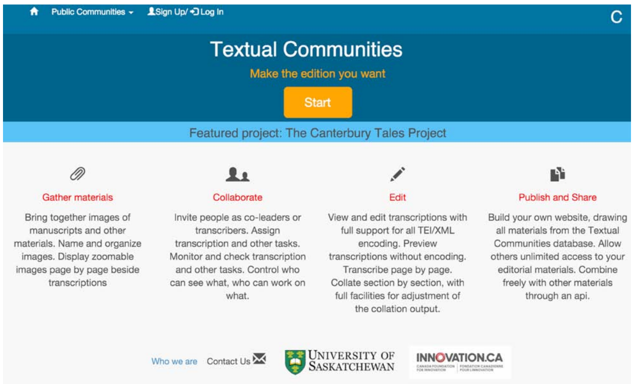
There are many editorial tools under development and a few that are already functional. There are two defining features that make Textual Communities different from the rest: its integrated participant and document management systems, and its mapping of fundamental document-entity structure. These features correspond to two underlying principles: that the work of amassing large corpora of textual materials is best accomplished by a well-managed community of interested participants from within, but also potentially from outside the academy; and that for the resulting materials to be useful, their relationships must be clearly articulated.

As its name suggests, Textual Communities is designed for gathering and organizing multiple participants around a common editorial project. It supports a wide range of relational structures, from a carefully crafted team to ad hoc community built on crowd-sourcing. Crucially, it enables definition of roles in the project with varying degrees of access to project materials and of authority to do the work of processing these materials, and oversight over other participants. It is also built on an ontology of text, document and communicative act to identify and relate the produced transcriptions ("texts"), the exemplars they derived from ("documents," usually in the form of a digital image of a particular witness), and the intellectual construct they instantiate (the communicative act or "work", or our preferred term, "entity"). Thus anyone interested in John Donne's poem "The Good Morrow" will find various "texts" (transcriptions) of this work as found in the extant "documents" (the poem as it is found in each of the manuscripts and printed books that contain it).

Textual Communities itself enables uploading of digital images of primary and linkage of these images with a transcription space. The user supplies information for each document, which defines the text that is to be transcribed and its relationship to the source document and entity. The user also defines the structure of the document, which is rendered behind the visible transcription in TEI conformant XML. The transcription area, which is automatically linked to the source image, can also support any XML markup that is desired or required for intelligent transcription of the source document. The collation area (based on CollateX) offers full regularization and editing facilities, to enable creation of optimal collations for scholarly use.

This open-source tool is available free of charge for use and adaptation by anyone anywhere. Development of this tool is funded by a generous grant from the Canadian Foundation for Innovation with the support of the Digital Research Centre at the University of Saskatchewan.

This poster will be accompanied by a live demonstration of the editorial workspace.



Public Communities ▾ Sign Up/ Log In

Textual Communities

Make the edition you want

Start

Featured project: The Canterbury Tales Project

Gather materials

Bring together images of manuscripts and other materials. Name and organize images. Display zoomable images page by page beside transcriptions

Collaborate

Invite people as co-leaders or transcribers. Assign transcription and other tasks. Monitor and check transcription and other tasks. Control who can see what, who can work on what.

Edit

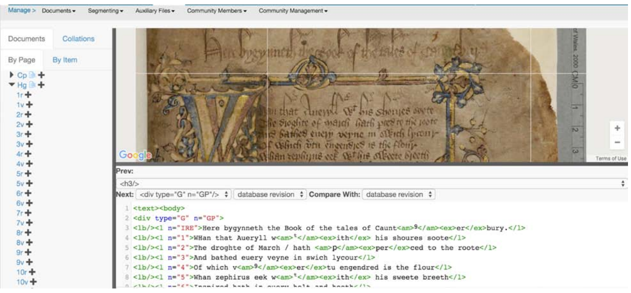
View and edit transcriptions with full support for all TEI/XML encoding. Preview transcriptions without encoding. Transcribe page by page. Collate section by section, with full facilities for adjustment of the collation output.

Publish and Share

Build your own website, drawing all materials from the Textual Communities database. Allow others unlimited access to your editorial materials. Combine freely with other materials through an api.

Who we are Contact Us

Figure 1: opening screen of Textual Communities



Manage > Documents ▾ Segmenting ▾ Auxiliary Files ▾ Community Members ▾ Community Management ▾
Documents
Collations
By Page
By Item
Cp
Hg
1r
1v
2r
2v
3r
3v
4r
4v
5r
5v
6r
6v
7r
7v
8r
8v
9r
9v
10r
10v

Prev: <h3/>

Next: <div type="G" n="GP"/> database revision Compare With: database revision

```
<text><body>
<div type="G" n="GP">
<lb/><l n="IRE">Here bygynneth the Book of the tales of Caunt<am>9</am><ex>er</ex>bury.</l>
<lb/><l n="1">WHan that Aueryll w<am>t</am><ex>ith</ex> his shoures soote</l>
<lb/><l n="2">The droghte of March / hath <am>p</am><ex>per</ex>ced to the roote</l>
<lb/><l n="3">And bathed euery veyne in swich lycour</l>
<lb/><l n="4">Of which v<am>9</am><ex>er</ex>tu engendred is the flour</l>
<lb/><l n="5">Whan zephirus eek w<am>t</am><ex>ith</ex> his sweete breeth</l>
```

Figure 2: transcription interface

Multivalent reuse of web data about temporary art exhibitions: the Exhibitium project

Nuria Rodríguez-Ortega
nro@uma.es
University of Málaga, Spain

José Pino Díaz
jospindia@uma.es
University of Málaga, Spain

Juan Luis Suárez
jsuarez@uwo.ca
University of Western Ontario, Canada

Rafael Bailón Moreno
bailonm@gmail.com
University of Granada, Spain

Our proposal aims to display the analysis techniques, methodologies as well as the most relevant results expected within the Exhibitium project framework (http://www.exhibitium.com). Awarded by the BBVA Foundation, the Exhibitium project is being developed by an international consortium of several research groups [1]. Its main purpose is to build a comprehensive and structured data repository about temporary art exhibitions, captured from the web, to make them useful and reusable in various domains through open and interoperable data systems.

The Exhibitium project aims to face the new challenges posed by the data-driven society respect to the production, management, use and distribution of digital cultural content. Specifically, the question that we want to respond is: how to take advantages of the universe of data related to cultural activity distributed through the Internet to generate value at very different levels (academic research, new critical narratives, cultural management, economic development, processes of social and cultural transformation, etc.)? Accordingly, multivalency, as a key notion of the new knowledge economy (Roney et al., 2012), is the guiding axis of the Exhibitium project.

To think on a precise working prototype, we have adopted as a central object of our research the art exhibitions that regularly held galleries, museums and art centers, since they produce a very large and heterogeneous set of data that can potentially be captured by means of web mining strategies.

For that, we have developed a technology infrastructure, which has been implemented in two dimensions:

a) A computing prototype consists of two modules operating in a complementary manner: Beagle, an automated data capture system which extracts information about art exhibitions from any web source; and ExpoFinder, a system for collecting and enhancing the information captured by Beagle (http://www.expofinder.es).

b. An aggregate of three data analysis and visualization platforms (SylvaDB, Techne Coword © and Geowave), which connect with the Beagle-ExpoFinder system using an automated export application [2]. (See in figure 1 a simplified diagram of the system).

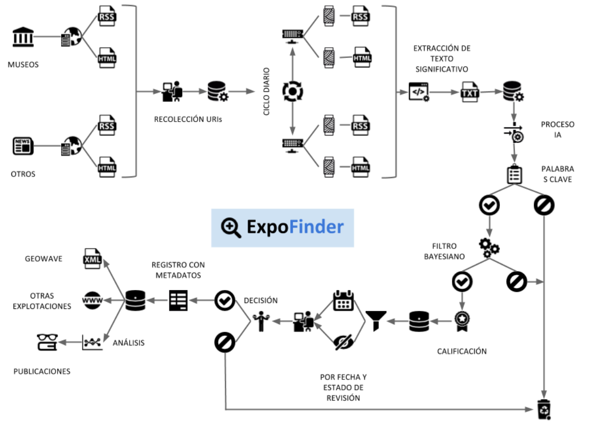


Figure 1. Simpiflied schema and flowchart

As regards the theoretical framework, Exhibitium faces the phenomenon of art exhibitions according to the actor-network theory (ANT) developed by Bruno Latour (2007), that is say, as complex cultural phenomena resulting from of a series of relationships established between heterogeneous human and no human actors -artists, curators, exhibition centers, funding agencies, publishing companies, artistic movements, artistic themes, etc.- that constitute dynamic networks among them. The application of the actor-network theory to the field of art exhibitions raises an interesting perspective, since the emphasis, rather than on the exhibitions themselves, is placed on the «mediation» processes (production, distribution, reception, etc.) as part of a social system in which the power relations that characterize the cultural institutions are unveiled.

Consequently, the Exhibitium project aligns with the growing research field oriented to explore complex networks in art and humanities (Schich, 2014). Particularly, we are using two types of methodologies.

a) Network analysis strategies through a graph database (SylvaDB) that we have structured according to an ontology specifically developed for the project (see figure 2).

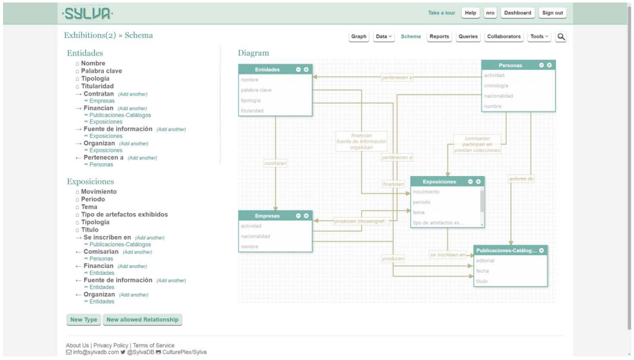


Figure 2. Ontology of the "temporary art exhibition" domain in the SylvaDB interface

b) KDD (Knowledge Discovery in Databases) techniques, specially the co-word analysis strategy (He, 1999) [3], which are mainly oriented to the discovery of implicit knowledge in large repositories of structured data (Echevarría and González, 2009; Rodríguez Ortega et al, 2015).

Through these analyses we are getting a set of graphs that is allowing us to discover correlations between certain artists, centers, curators, financing institutions, themes and geo-spatial contexts. Note that the configuration of clusters of this nature, understood as network of strong relationships established between their nodes, can be read as power structures operant in a certain domain.

The conceptual and methodological differences existing between the co-word analysis strategy, which extracts the actors and their relationships directly from the data processing, and the network analysis strategy, based on a scheme previously established (ontology) that derives from our direct observation of the real world, it is allowing us to make interesting comparisons between how we understand the domain "art exhibitions" and what (implicit) structure emerges naturally from the processing of the keywords associated to this field.

Along with this knowledge -key to develop new critical narratives-, the discovery of certain patterns is also providing a prospective knowledge about possible and future trends in the field of temporary art exhibitions, which can be strategically used by exhibition centers or other stakeholders (tourism companies, entertainment industries, etc.) in their decision-making processes.

In addition, these methods of analysis are allowing us to investigate the peripheries that are also associated to the field of art exhibitions as a result of the prevalence of some institutions which centralize the flow of public among other things due to their media impact capacity. To facilitate the social visibility of those initiatives usually unknown we are proceeding to integrate the Exhibitium data in the Geowave system, a georeferenced platform to manage cultural content.

Notes

[1] They are: iArtHis_Lab (http://www.iarthislab.es) and Khaos (http://khaos.uma.es) at the University of Málaga; Techne, ingeniería del conocimiento y del producto (http://www.ugr.es/~tepo28/quienes_somos_es.php) at the University of Granada; and CulturePlex at the University of Western Ontario (http://www.cultureplex.ca).

[2] Since the nature of this text does not allow us to provide a detailed explanation of such system, we suggest to see the site http://admin.expofinder.es/test/xmlrpc/ [Viewed: 28/10/2015].

[3] The co-word strategy is based on a basic principle: all actors in a network, regardless of their nature, can be represented by words or descriptors (actually, this is how the data are represented mostly in bibliographic repositories). Natural relations of co-occurrence of words in data repositories form a network that can be analyzed and mapped. The words that co-occur more frequently form clusters that can be analyzed as networks with strong links; at the same time, it is possible to analyze how these networks change throughout different periods of times or geo-spatial contexts. For example, we can analyze how the network generated by an artist (i.e., Picasso) changes according to the geographical provenance of the data (exhibitions in Spain, Europe, USA, Latin America, etc.).

Bibliography

Echevarría, J. and González, M. I. (2009). La Teoría del Actor-Red y la tesis de la Tecnociencia, *Arbor Ciencia, Pensamiento y Cultura*, **738**: 705-20.

He, Q. (1999). Knowledge Discovery Through Co-Word Analysis, *Library Trends*, **48**(1): 133-59.

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory.* Oxford and New York: Oxford University Press.

Rodríguez Ortega, N. et al. (2015). Repesando los estudios metadisciplinares en la sociedad datacéntrica: análisis dinámico de las redes de conocimiento de la Historia del Arte a través de la base de datos ISOC-Arte del CSIC. *II Congreso de la Sociedad Internacional de Humanidades Digitales Hispánicas.*

Innovación, globalización e impacto, 5-7 de octubre de 2015, Madrid: UNED-Createspace, pp. 322-36.

Roney, D. et al. (2012). Knowledge is people doing things, knowledge economies are people doing things with better outcomes for more people. In *Handbook on the Knowledge Economy*, Elgar Original Reference Series, Edward Elgar Publishing, Incorporated, pp. 1-14.

Schich, M. et al. (2014). *Arts, Humanities, and Complex Networks*, 4th edition, *Leonardo,* Cambridge: MIT Press.

Critical Edition as Graph: The Chronicle of Matthew of Edessa Online

Anahit Safaryan
anahit.safaryan@students.unibe.ch
Universität Bern, Switzerland

Sascha Kaufmann
sascha.kaufmann@dh.unibe.ch
Universität Bern, Switzerland

Tara Lee Andrews
tara.andrews@kps.unibe.ch
Universität Bern, Switzerland

In the last few years, several of the bodies that govern higher education and research in Switzerland – the swissuniversities council, the Swiss National Science Foundation (SNF), and the Swiss Academy of Humanities and Social Sciences (SAGW) – have placed an increasing focus on the importance of appropriate and sustainable handling research data, and in particular research data in the humanities. As ever more research in the humanities includes collections of digital data, these Swiss bodies are putting a significant investment into building an infrastructure capable of maintaining it. Against this background, the SNF has begun to focus in particular on the scholarly text edition; beginning in 2017, all SNF-funded edition projects are expected to be digital and to contain some strategy for long-term preservation (Swiss National Science Foundation, 2015). One of the purposes of the SNF-funded project "The Chronicle of Matthew of Edessa Online" is to experiment with forms of digital critical edition of texts that can be preserved and maintained in the long term.

The Chronicle is an Armenian-language historical work, written in the early 12th century, covering nearly two centuries' worth of the history of the medieval Near East, up to and including the establishment of the Crusader County of Edessa. The text survives in at least 35 manuscript copies, all from 1590 or later. Digital methods have been used since the beginning of the edition work in 2008; this includes full transcriptions of the individual manuscripts in a TEI-compatible markup system, automatic collation of the textual variants, computer-assisted stemmatic analysis, and systematic editorial analysis of the variant texts (Andrews, 2009). Since the work began, however, there has been significant progress in thought about how to represent textual variants in digital space, most particularly with the adoption of the variant graph (Andrews, 2014; Dekker, Hulle, Middell, Neyt, and Zundert, 2014; Jänicke, Geßner, Büchler, and Scheuermann, 2014; Schmidt and Colomb, 2009).

In this poster session we will demonstrate the ways in which, beginning with the variant graph concept, we are implementing all aspects of the edition in a graph database model – the textual collation itself, analysis of the manuscripts and stemmata, the identification of events, timelines, persons, and places that normally occupies much of the commentary of the edition, and the annotations that together with these identifications will comprise the full commentary. Since it is a graph, the data becomes straightforward to transform directly into the Resource Description Framework (RDF) required for compatibility with the SALSAH database provided by the SAGW-funded Data and Service Centre for the Humanities, which is the nascent repository for all humanities research data within Switzerland and guarantees long-term API-addressable accessibility to the data itself (Rosenthaler, Fornaro, and Clivaz, 2014). Moreover, in the construction of the graph model we draw upon relevant sets of standards and tools, such as the TEI Guidelines (http://www.tei-c.org/) for description of the textual features, CIDOC-CRM (http://www.cidoc-crm.org/) to express information about the documents as objects and for notations of time and place in a manner suitable to use with tools such as Topotime (Grossner and Meeks, 2014). We are also participating in the DARIAH-DE Working Group on Digital Annotations in order to ensure that our working methods are methodologically and practically compatible with best practice in humanities research.

The result of the project will be a digital edition whose constructed text, scholarly commentary, and individual manuscript witnesses will be fully accessible through a machine-queryable API. The web user interface will be built entirely on this API; in that way, no intellectual content will be embedded in the interface, only to be eventually lost when or if the interface disappears. The manuscript texts and a version of the edition itself will be available to download in TEI and ePub formats, for print consumption, archiving, or further offline analysis.

Bibliography

Andrews, T. L. (2009). Prolegomena to a Critical Edition of the Chronicle of Matthew of Edessa, with a Discussion of Computer-Aided Methods Used to Edit the Text. University of

Oxford, http://ora.ouls.ox.ac.uk/objects/uuid%3A67ea947c-e3fc-4363-a289-c345e61eb2eb.

Andrews, T. L. (2014). Analysis of Variation Significance in Artificial Traditions Using Stemmaweb. *Digital Scholarship in the Humanities*, doi: 10.1093/llc/fqu072.

Dekker, R. H., Hulle, D. van, Middell, G. et al. (2014). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Literary and Linguistic Computing*, fqu007. 10.1093/llc/fqu007.

Grossner, K. and Meeks, E. (2014). Topotime: Representing historical temporality. *Proceedings of DH2014*, Lusanne.

Jänicke, S., Geßner, A., Büchler, M. et al. (2014). 5 Design Rules for Visualizing Text Variant Graphs. Presented at the Digital Humanities 2014, Lausanne, http://dharchive.org/paper/DH2014/Paper-652.xml (accessed 27 October 2015).

Rosenthaler, L., Fornaro, P. and Clivaz, C. (2014). Long term preservation of digital information. Presented at the Digital Humanities 2014, Lausanne, http://dharchive.org/paper/DH2014/Paper-68.xml (accessed 27 October 2015).

Schmidt, D. and Colomb, R. (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, **67**: 497–514.

Swiss National Science Foundation. (2014), Call für Editionsprojekte mit Blick auf die Finanzierungsperiode 2017–2010, http://www.snf.ch/SiteCollectionDocuments/call_editionen_phase2_d.pdf (accessed 27 October 2015).

Development of a Support Tool for Categorizing Ukiyo-e's Pictorial Themes: A System to Deal with Visual Features and Similarities

Shinya Saito
saitos@fc.ritsumei.ac.jp
Ritsumeikan University, Japan

Keiko Suzuki
suzukik@fc.ritsumei.ac.jp
Ritsumeikan University, Japan

I. Introduction

This research aims at supporting categorization of ukiyo-e (Japanese woodblock prints) by developing an original, data-visualization system. As a part of the system development, this case study focuses on the prints' visual features—how the system can help analyzing them by simulating the analytical process. Our investigation of the ukiyo-e research's process helps us identifying what the researchers actually need to know and how to deal with it.

II. Research process

For developing the system, we take an example of pictorial theme of "Otohime," who is these days explained as a mythical princess, living in Ryugu or the Dragon Palace at the bottom of the sea. Princess Otohime or comparable others related to the Dragon Palace have appeared in many stories with versions and variants through history, whose prolificness their ukiyo-e imagery reflect. Thus, we need to pay our attention to this kind of pictorial themes' extensibility in intertextuality, or in this case, inter-imagery, by which one image refers to, exploits, or recycles another. For comparing and analyzing the heroin's visual features thoroughly, the following 9 visual features of hers are chosen as indexes: 1. headdress; 2. hairstyle; 3. frills; 4. scarf; 5. Chinese fan; 6. Chinese-style clothes; 7. collar; 8. apron; and 9. Urashima. Checking the indexes leads researchers to understand the degree of similarity in different ukiyo-e prints, and generate hypotheses about what kinds of factors affect specific ukiyo-e production as well as historical changes in Otohime's overall imagery.

III. System development

Recently, data visualization, both its technology applicable for many fields and methodology, has been systematically developed (Mazza and Berre, 2007; Tufte, 1983). In order to pursue the above-mentioned purpose, we are developing the SALOMONIS system (Fig.1) which can load an ukiyo-e dataset encoded in JSON (JavaScript Object Notation) format. The following list shows how SALOMONIS visualizes the data.

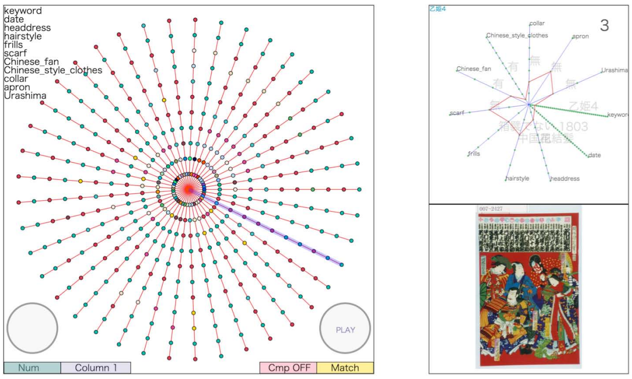


Fig. 1. Layout chart of SALOMONIS

A. Entire dataset and records

In the system, a record is visualized as a line, each of which is arranged in a radial manner. As a result, SALOMONIS visualizes the records, i.e., the entire dataset, as a circle (Fig.2).

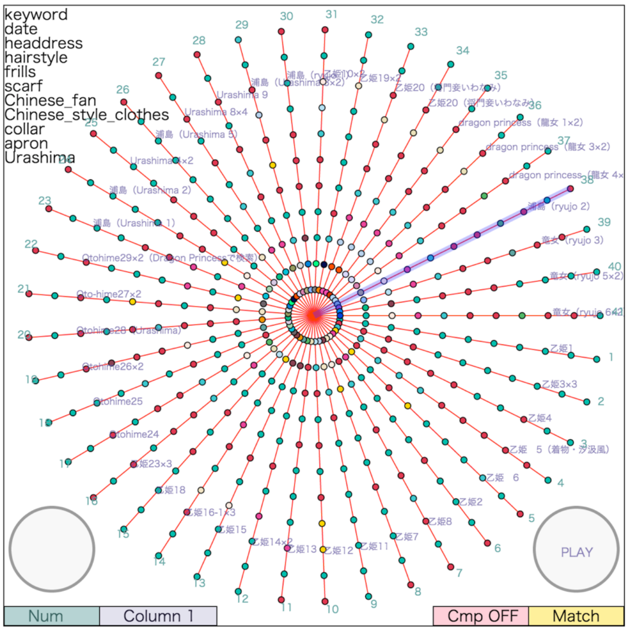


Fig. 2. Visualization of entire dataset

B. Columns

As mentioned above, a line indicates a record, and each record includes columns that correspond to indexes. Our system visualizes columns as dots, plotted on the line. Each color of the dots indicates different value of the column.

In the case of the Otohime dataset, it has 41 records and 11 columns with 9 visual features and the print's production year and keyword. Therefore, 41 lines are arranged, each of which has 11 dots (Fig.3).

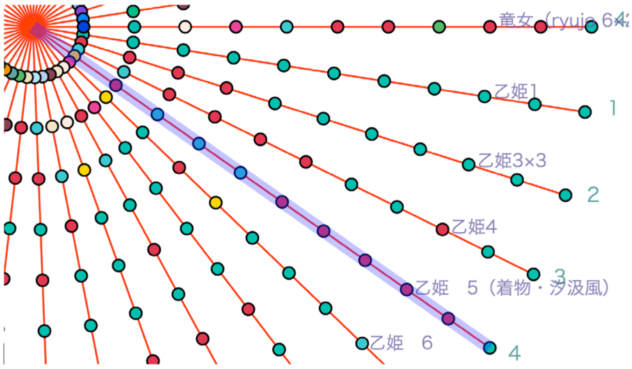
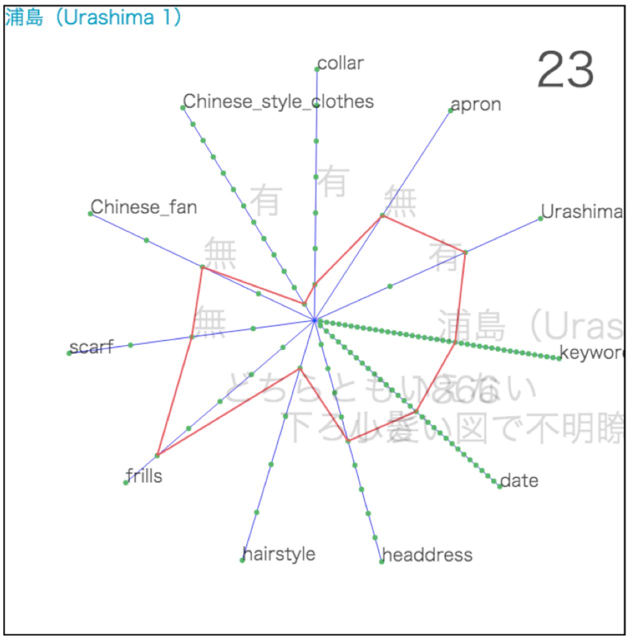


Fig. 3. Visualization of columns

C. Interactive function

As each line is linked to an ukiyo-e print, when the user puts the mouse cursor on a certain line, a chart appears with its corresponding ukiyo-e's 9 visual features plus its production year and keyword, mentioned on the above (Fig.4). Using this interactive function, the user can compare ukiyo-e in a speedy and accurate manner.



>Fig. 4. Chart that indicates ukiyo-e's features

D. Similarity-screening function

When the user chooses one ukiyo-e as a reference point, then pushing the "Play" button, the system starts automatically screening how the other ukiyo-e are similar or not to the chosen one. Depending upon the degree of the similarity, lines of the ukiyo-e appear highlighted with colors. For example, when one ukiyo-e shares 80 percent or more similarity with the chosen one in terms of the visual features, the line appears in orange. In the case of 60 percent or more, green; 40 percent or more, yellow; and 20 percent or more, red (Fig.5).

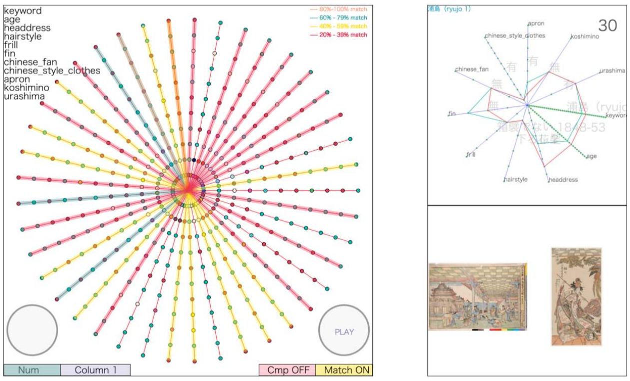


Fig. 5. Similarity-screening function

E. Visual Sort function

Furthermore, this system has visual sort function. Using this, all records are sorted based on the similarity in a radial manner.

IV. Discussion

The system proved its efficiency as it not only saved the time for the analyses but also suggests new ways to think about the prints. Right now, however, we are still in the developing stage, planning to test the system by applying it for more cases.

Bibliography

Mazza, R. and Berre, A. (2007). Focus group methodology for evaluating information visualization techniques and tools, *Proceedings of the 11th IEEE International Conference on Information Visualisation*, pp. 74-80.

Tufte, E. (1983). *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

eComparatio - Editionsvergleich

Charlotte Schubert
schubert@uni-leipzig.de
Universität Leipzig, Germany

Friedrich Meins
friedrich_meins@uni-leipzig.de
Universität Leipzig, Germany

Oliver Bräckel
oliver.braeckel@uni-leipzig.de
Universität Leipzig, Germany

Hannes Kahl
hannes.kahl@uni-leipzig.de
Universität Leipzig, Germany

Das von der Deutschen Forschungsgemeinschaft (DFG) geförderte Projekt eComparatio wird seit 2014 als Kooperationsprojekt des Lehrstuhls für Alte Geschichte der Universität Leipzig und des ICE (Interdisciplinary Center of E-Humanities in History and Social Sciences/ Forschungsstelle am Max-Weber-Kolleg für kultur- und sozialwissenschaftliche Studien an der Universität Erfurt) entwickelt. Das Kernstück der Anwendung ist ein Modul zum Vergleich von Textausgaben, das auch die Erstellung eines Variantenapparates für digitale Editionen antiker Autoren ermöglicht. Die Zahl der Vergleichstexte ist beliebig, ebenso das Eingabeformat (TXT, HTML, XML, JSON, PDF). Die Anwendung wird frei skalierbar sein, so dass der Umfang der zu vergleichenden Texte nicht beschränkt ist. Das Ergebnis (Kollationierung) soll in Form von Listen als Apparat (positiver oder negativer Apparat) oder auch in beliebiger anderer Form ausgegeben werden können. In einem weiteren Modul ist für Autorenreferenzen bei der Abfrage von online-Datenbanken die Anbindung an das Referenzsystem CTS (Canonical Text Services) ermöglicht worden. Im bisherigen Verlauf des Projektes ist es gelungen, die Grundfunktionen des Tools zu implementieren und es in die Lage zu versetzen eine beliebig große Anzahl an Texten miteinander zu vergleichen. Dabei sind drei unterschiedliche Ansichten entstanden, die es dem Benutzer ermöglichen, das Ergebnis aus verschiedenen Perspektiven zu betrachten. Die Detailansicht zeigt einen Text und markiert entsprechende Unterschiede zu anderen Texten. Die Parallelansicht (siehe auch Abbildung) zeigt alle Texte nebeneinander und markiert die Unterschiede farbig. Die Buchansicht schließlich zeigt wieder nur einen Text an und visualisiert die Varianten im Stile traditioneller Printeditionen unter dem betreffenden Abschnitt. Zu betonen ist dabei, dass der Ausgangstext für den Vergleich bei jeder Ansicht frei wählbar ist und sich somit nicht auf einen zu bevorzugenden Haupttext festgelegt bzw. eine Gewichtung der Textzeugen vorgenommen wird.

Die Visualisierung und Ergebnissicherung ermöglicht zum einen, einen schnellen Überblick über die Text- und Editionsgeschichte verschiedener in digitalisierter Form vorliegender Werke zu erlangen. Darüber hinaus eignet sich das Tool als Hilfsmittel zum Kollationieren bei der Erstellung beliebiger kritischer, historischer bzw. genetischer Editionen.

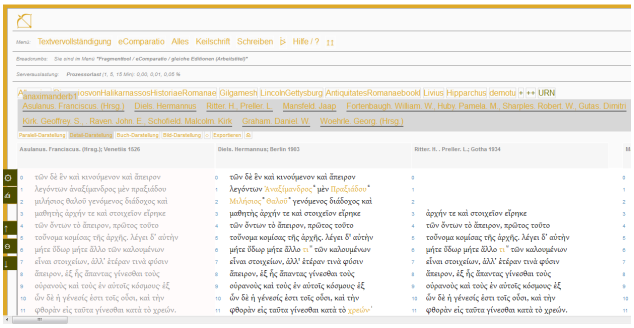
Im Rahmen zweier Folgeprojekte in Kooperation mit Prof. Christopher Blackwell von der Furman University in Greenville, SC, von denen eines sich derzeit (Stand: Oktober 2015) in der Antragsphase (DFG) befindet und das andere bereits bewilligt worden ist (Andrew W. Mellon-Foundation), wird zunächst auf Basis einer Kooperation mit dem De Gruyter-Verlag und der durch ihn bereitgestellten digitalen Ausgabe der Bibliotheca Teubneriana Latina Online die Basis für eine kanonische CTS-Referenzierung geschaffen. Auf dieser Grundlage wird das eComparatio-Vergleichstool dazu dienen, sämtliche digital zur Verfügung stehenden Exemplare eines Textes durch eine CTS-Anfrage zu vergleichen, zu visualisieren und für die Ergebnisse die jeweiligen Textpassagen auch mit einer maschinenlesbaren Referenzierung zu versehen. Ein solches Verfahren kann dabei als wesentlicher Schritt der Qualitätssicherung der digitalen Datengrundlage an sich angesehen werden. Gerade im Falle der Altertumswissenschaften, in denen bereits früh umfangreiche, abgeschlossene Korpora (TLG, BTL u.a.) vorlagen, ist ein nächster Schritt ein Ausbau dieser Datengrundlagen in die Tiefe, d.h. hinsichtlich der zahlreichen verschiedenen Editionen und Textausgaben.

Darüber hinaus wird in einem USE-Case am Beispiel der Periklesvita des Plutarch in Verbindung mit der Zitationsanalyse aus dem von 2008-2013 durch das BMBF geförderten Projekts eAQUA des Lehrstuhls für Alte Geschichte der Universität Leipzig ein praktisches Beispiel gegeben, inwiefern CITE/CTS als Grundlage einer neuen Form des Annotierens und Kommentierens dienen kann:

Der Zitatonsgraph dient zur automatischen Erstellung eines Testimonien-, eComparatio zur Erstellung eines Variantenapparates. Das Verhältnis der verschiedenen Texte zueinander (Varianten des Haupttextes und der enthaltenen Fragmente/Zitate/Parallelstellen) soll darüber hinaus durch die im Rahmen des Projektes von den amerikanischen Projektpartnern entwickelten Erweiterung Canonical Graph Services (CGS) visualisiert werden.

Nach seiner Fertigstellung wird das Tool als freier Webservice für Forschung und Lehre zur Verfügung gestellt. Davon können Handschriften-Digitalisierungsprojekte, Editionsprojekte sowie Projekte profitieren, die sich Spezialfragen einzelner Textpassagen widmen; es ist auch für Seminararbeiten, d.h. den Einsatz in der Lehre geeignet, da es sowohl von Nicht-Editionsphilologen als auch von Editionsphilologen eingesetzt werden kann.

Da es sich bei dem Tool in erster Linie um ein Mittel zur Visualisierung handelt, ist es in hohem Maße für die Präsentation in Form eines Posters geeignet. Geplant ist die Darstellung des gesamten Workflows anhand eines Beispiels, von der Eingabe unstrukturierter Textdokumente bis hin zu den drei oben genannten Visualisierungsformen sowie erster Ergebnisse der begonnenen Folgeprojekte.


Menü: Textvervollständigung eComparatio Alles Keilschrift Schreiben Hilfe / ? ↿⇂
anaximanderb1 DionysiosvonHalikarnassosHistoriaeRomanae Gilgamesh LincolnGettysburg AntiquitatesRomanaebookI Livius Hipparchus demotu + ++ URN
Asulanus. Franciscus. (Hrsg.) Diels. Hermannus Ritter. H., Preller. L. Mansfeld. Jaap Fortenbaugh. William. W., Huby. Pamela. M., Sharples. Robert. W., Gutas. Dimitri
Kirk. Geoffrey. S., . Raven. John. E., Schofield. Malcolm. Kirk Graham. Daniel. W. Woehrle. Georg. (Hrsg.)
Paralell-Darstellung Detail-Darstellung Buch-Darstellung Bild-Darstellung ○ Exportieren

Breadcrumbs: Sie sind im Menü "Fragmenttool / eComparatio / gleiche Editionen (Arbeitstitel)"

Serverauslastung: Prozessorlast (1, 5, 15 Min): 0,00, 0,01, 0,05 %

	Asulanus. Franciscus. (Hrsg.); Venetiis 1526		Diels. Hermannus; Berlin 1903		Ritter. H. . Preller. L.; Gotha 1934	M
0	τῶν δὲ ἓν καὶ κινούμενον καὶ ἄπειρον	0	τῶν δὲ ἓν καὶ κινούμενον καὶ ἄπειρον	0		0
1	λεγόντων ἀναξίμανδρος μὲν πραξιάδου	1	λεγόντων Ἀναξίμανδρος μὲν Πραξιάδου	1		1
2	μιλήσιος θαλοῦ γενόμενος διάδοχος καὶ	2	Μιλήσιος Θαλοῦ γενόμενος διάδοχος καὶ	2		2
3	μαθητὴς ἀρχήν τε καὶ στοιχεῖον εἴρηκε	3	μαθητὴς ἀρχήν τε καὶ στοιχεῖον εἴρηκε	3	ἀρχήν τε καὶ στοιχεῖον εἴρηκε	3
4	τῶν ὄντων τὸ ἄπειρον, πρῶτος τοῦτο	4	τῶν ὄντων τὸ ἄπειρον, πρῶτος τοῦτο	4	τῶν ὄντων τὸ ἄπειρον, πρῶτος τοῦτο	4
5	τοὔνομα κομίσας τῆς ἀρχῆς. λέγει δ' αὐτὴν	5	τοὔνομα κομίσας τῆς ἀρχῆς. λέγει δ' αὐτὴν	5	τοὔνομα κομίσας τῆς ἀρχῆς. λέγει δ' αὐτὴν	5
6	μήτε ὕδωρ μήτε ἄλλο τῶν καλουμένων	6	μήτε ὕδωρ μήτε ἄλλο τι τῶν καλουμένων	6	μήτε ὕδωρ μήτε ἄλλο τι τῶν καλουμένων	6
7	εἶναι στοιχείων, ἀλλ' ἑτέραν τινὰ φύσιν	7	εἶναι στοιχείων, ἀλλ' ἑτέραν τινὰ φύσιν	7	εἶναι στοιχείων, ἀλλ' ἑτέραν τινὰ φύσιν	7
8	ἄπειρον, ἐξ ἧς ἅπαντας γίνεσθαι τοὺς	8	ἄπειρον, ἐξ ἧς ἅπαντας γίνεσθαι τοὺς	8	ἄπειρον, ἐξ ἧς ἅπαντας γίνεσθαι τοὺς	8
9	οὐρανοὺς καὶ τοὺς ἐν αὐτοῖς κόσμους· ἐξ	9	οὐρανοὺς καὶ τοὺς ἐν αὐτοῖς κόσμους· ἐξ	9	οὐρανοὺς καὶ τοὺς ἐν αὐτοῖς κόσμους· ἐξ	9
10	ὧν δὲ ἡ γένεσίς ἐστι τοῖς οὖσι, καὶ τὴν	10	ὧν δὲ ἡ γένεσίς ἐστι τοῖς οὖσι, καὶ τὴν	10	ὧν δὲ ἡ γένεσίς ἐστι τοῖς οὖσι, καὶ τὴν	10
11	φθορὰν εἰς ταῦτα γίνεσθαι κατὰ τὸ χρεών.	11	φθορὰν εἰς ταῦτα γίνεσθαι κατὰ τὸ χρεών·	11	φθορὰν εἰς ταῦτα γίνεσθαι κατὰ τὸ χρεών.	11

Abb. der Parallelansicht von eComparatio am Beispiel des Fragments B1 des Anaximander.

Authorship Attribution of Mediaeval German Text: Style and Contents in Apollonius von Tyrland

Sarah Schulz
sarah.schulz@ims.uni-stuttgart.de
University of Stuttgart, Germany

Jonas Kuhn
jonas.kuhn@ims.uni-stuttgart.de
University of Stuttgart, Germany

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
University of Stuttgart, Germany

Introduction

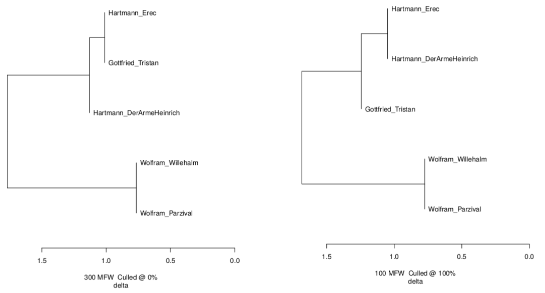
In this paper, we describe computer-aided authorship testing on the Middle High German (MHG) text *Apollonius von Tyrland* written by Heinrich von Neustadt (HvN) in the late 13th century. Being based on a Latin original, HvN is suspected to incorporate other sources into the translation. We investigate assumptions regarding a segmentation of this text into parts supposedly tracking back to different sources. Our objective is it to provide a) clarification on the validity of this segmentation and b) on features that show the difference in origin of the segments. In particular, we distinguish between features related to content and to style.

Contents and Style

Comparing frequency distributions over frequent words has been established as a state of the art method for contrasting style across different literary texts (cf. Eder et al. (2013)). Quite recently, (Herrmann et al., 2015) proposed to define style as a property constituted by "formal features which can be observed quantitatively or qualitatively" (p. 44). An important aspect of it is that style has to be based on observable features.

We propose to make a clear cut between content and style: To measure stylistic differences, we restrict the selection to words appearing in every text of the corpus, thus are observable in each text, assuming that this is a simple way to exclude words that are markers of content. Content words (that presumably only appear in a subset of the texts) do not contribute to this understanding of style. They, in contrast, are extracted by filtering the MFW with a stop word list containing all the function words in a language. We refer to the sets of feature words extracted for a text with **content words** and **style words**.

To validate this idea, we analyse five MHG texts by three authors with the R stylo package (Eder, 2013). Figure 1 shows results for the content (a) and style (b) words. The higher similarity of *Erec* and *Tristan* in (a) compared to *Der arme Heinrich* reflects that both narratives feature knighthood as a main theme. In contrast, the narrative in *Der Arme Heinrich* involves more religious themes (faith, god), which is also reflected in the frequency tables. This distinction is clearly based on content. If we focus our analysis on style words, as in (b), we see the clustering according to authors. Thus, distinguishing frequent words of a corpus in style and content words can give us better insights into the results.



(a) Clustering analysis using content words.

(b) Clustering analysis using style words.

Figure 1: Comparison of different groups of high frequent words and their performance on a clustering task on MHG text by three different authors. Due to largely uniform editing of MHG text in the 19th century, normalisation can be neglected (Kragl, 2015).

Dissecting Heinrich von Neustadt: Apollonius

Nr.	Verses	Origin
1	1-2,905	Latin original
2	2,906-15,118	Insertion
3	15,119-17,382	Latin original
4	17,383-20,589	Insertion

Table 1: Partition of Apollonius according to Bockhoff and Singer (1911).

Nr.	Verses	Title
1	2920-4126	The fight with Gog and Magog
2	4126-6068	The adventures in Galacites
3	6069-7186	The duel in Syria and the Robinson Island
4	7187-10594	Bulgare war and imprisonment in Nemrot
5	10595-13512	The adventures in Chrysa
6	13517-14929	The return to Tarsus
7	17282-20639	Closing

Table 2: Subparts identified in the third section of Apollonius, identified by Bockhoff and Singer (1911).

Bockhoff and Singer (1911) formulated two hypotheses regarding the internal structure and origin of the ca. 21K verses long text, regarding both the overall structure (Table 1) and the internal structure of one segment (Table 2). To get an impression of which paragraphs can indeed be found as a distinctive group using content words and style words respectively, we split the text into 71 segments of equal length. These segments are then clustered with Stylo, using delta as a similarity measure.

Our baseline consists of randomly assigning distances between the segments, drawn from a uniform distribution. We sample baseline results 1000 times. The baseline results give an impression on how well an uninformed method overlaps with the hypothesis classification introduced in Section 3. Comparing these to the results of our methods can inform us on which method goes in line with the hypothesis as opposed to a random classification.

Features	Part	Recall		Precision		F-score	
		BL	CA	BL	CA	BL	CA
Content words	Historia Apollonii 1	0.46	0.56	0.19	0.4	0.19	0.47
	Adventure	0.33	0.23	0.66	1.0	0.44	0.37
	Historia Apollonii 2	0.30	0.88	0.16	0.44	0.21	0.59
	Final apotheosis	0.19	0.45	0.17	0.25	0.18	0.32
Style words	Historia Apollonii 1	0.46	0.56	0.19	0.21	0.19	0.31
	Adventure	0.33	0.28	0.66	0.92	0.44	0.43
	Historia Apollonii 2	0.30	0.75	0.16	0.33	0.21	0.46
	Final apotheosis	0.19	0.72	0.17	0.5	0.18	0.67

Table 3: Results of the clustering analysis for style and content words respectively regarding the overall structure hypothesis of Apollonius. Since clustering methods do not provide class labels for an evaluation of the performance with respect to precision, recall and F-score, we need to map the clusters onto the parts of the hypothesis. This is done manually in such a way that F-score is maximised. BL: Baseline, CA: Cluster Analysis.

Regarding the first hypothesis (Table 3), we observe that for both feature sets the F-score lies above the baseline for all parts except the third. This seems reasonable since this part is suspected to be based on different sources and therefore might be more heterogeneous both in content and in style. Style seems to be more homogeneous (F-Score above baseline) throughout the entire text whereas content seems to be heterogeneous especially in the adventure part introduced by HvN (F-Score below baseline). This is in line with the hypothesis, considering that HvN's insertions report on different adventures.

Features	Part	Recall		Precision		F-score	
		BL	CA	BL	CA	BL	CA
content words	The fight with Gog and Magog	0.43	0.75	0.26	0.6	0.32	0.67
	The adventures in Galacites	0.36	0.50	0.32	0.5	0.34	0.5
	The duel in Syria	0.39	0.25	0.25	0.11	0.31	0.15
	Bulgare war	0.26	0.33	0.48	0.8	0.33	0.46
	The adventures in Chrysa	0.24	0.5	0.43	0.71	0.31	0.59
	The return to Tarsus	0.25	0.83	0.3	0.71	0.29	0.77
style words	The fight with Gog and Magog	0.43	0.75	0.26	0.375	0.32	0.5
	The adventures in Galacites	0.36	0.25	0.32	0.25	0.34	0.25
	The duel in Syria	0.39	0.5	0.25	0.33	0.31	0.4
	Bulgare war	0.26	0.5	0.48	0.46	0.33	0.48
	The adventures in Chrysa	0.24	0.4	0.43	0.67	0.31	0.5
	The return to Tarsus	0.25	0.33	0.3	0.4	0.29	0.36

Table 4: Results of the clustering analysis for style and content words respectively regarding structure of the parts of Apollonius attributed to Heinrich von Neustadt. Final part has been removed from the discussion due to its short length.

Analysing these heterogeneous parts further (second hypothesis, Table 4), we see heterogeneity in terms of content for all but one part, *The duel in Syria*. *The duel in Syria* seems homogeneous in style whereas *The adventures in Galacites* shows tendencies towards a heterogeneous style.

Conclusion

Both feature sets show similar tendencies and support a major part of the hypotheses by Bockhoff and Singer (1911) regarding parts suspected as insertions. Nevertheless, differences in content cannot clearly confirm the suspicion that HvN incorporated other sources. He might have created additional adventures by himself. Bockhoff and Singer (1911) do not cite sources from which HvN copied narratives, making it difficult to tackle exactly. Overall differences in style are much less significant than differences in content, which is in line with the hypothesis.

Bibliography

Bockhoff, A. and Singer, S. (1911). Heinrichs von Neustadt Apollonius von Tyrland und seine Quellen. Ein Beitrag zur mittelhochdeutschen und byzantinischen Literaturgeschichte von A. Bockhoff und S. Singer. *Sprache und Dichtung: Forschungen zur Linguistik und Literaturwissenschaft [dann] zur Sprach- und Literaturwissenschaft*. J. C. B. Mohr.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a Suite of Tools. In *Digital Humanities 2013: Conference Abstracts*, Lincoln, NE: University of Nebraska–Lincoln, pp. 487–89.

Eder, M. (2013). Mind your corpus: systematic errors in authorship attribution. *LLC*, **28**(4): 603–14.

Kragl, F. (2015). Normalmittelhochdeutsch. Theorieentwurf einer gelebten Praxis. *Zeitschrift fur Deutsches Altertum und Deutsche Literatur*, **144**: 1–27.

Herrmann, J. B., van Dalen-Oskam, K. and Schöch, C. (2015). Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory*, **9**(1): 25-52.

You Better Monetize! Monetization Strategies in Publishing and Disseminating Digital Scholarly Editions

Anna-Maria Sichani
anna-maria.sichani@huygens.knaw.nl
Huygens ING- KNAW / DiXiT, Netherlands, The

Aodhán Kelly
aodhan.kelly@uantwerpen.be
University of Antwerp / DiXiT

The digital scholarly edition is an academic product that attempts to remediate the theory and practice of print-based textual scholarship with computational methodologies. Although a vast amount of scholarly exertion is directed towards the exploration of digital editing as a scholarly enterprise (concepts, standards, technologies), significant discussions of sustainable models for publishing and disseminating the digital edition are rather infrequent. Such a practice could be explained mainly because digital editions are often developed as multi-institutional projects based on short-period funding and often every development and financial planning stops by the end of the funding period. Furthermore, as publication and distribution were procedures originally carried out by publishing houses, scholars and editors now find themselves ill-equipped and unprepared in coordinating such strategies. On the other hand, Open Access rhetoric advocating for "freely used, modified, and shared" scholarly content, put constantly into question principles of revenue generation strategies and their application.

In digital scholarship, there is obviously a cost for creating and maintaining a scholarly resource in the long-term. If these costs are not covered by a grant or by the institution, then the project must design and implement some sort of financial plan to cover it. This cost is not always treated as a price covered by the demand side (audience) or as a means to make profit - unlike in the world of commerce and publishing. It might, thus, be useful to re-think in what ways 'price' and 'value' could become interconnected in the web era. Such discussions around economic models were evident in the GLAM sector since the advent of the web, but have yet to become a major discussion in scholarly editing circles.

We would like to situate our exploratory approach within a broader discussion that challenges our traditional ideas towards the completion, publication, dissemination and sustainability of digital editions. Specifically, we will focus on economic models or monetization strategies as integral parts of a sustainability plan while embracing a strong Open Access ethos. How can we enrich the still burgeoning field of digital editing by contemplating the commodification and sustainability of our published scholarly outputs, by exploring current practices used in web publishing and e-commerce?

This research attempts to offer a state-of-the-art of current trends and business models strategies of distribution of digital content: subscription models and related price discrimination and audience segmentation principles, freemium and lite versions, referral discounts, value-added and tiered services, digital marketing, online advertising, derivatives, multimodal and customised formats for diverse purposes and audiences, etc. Our approach proposes a modular - in a "mix 'n' match" style - strategy for the monetisation of scholarly editions for their post-publication life, in which agents of digital editing projects could choose, combine and customize monetisation solutions for their digital scholarly editions in order to secure their sustainable future and impact in the long-term.

Bibliography

Eve, M. P. (2014). *Open Access and the Humanities: Contexts Controversies and Futures.* Cambridge: Cambridge University Press.

Greetham, D. C. (1994). *Textual Scholarship: An Introduction.* New York: Garland.

Greetham, D. C. (Ed.). (1995). *Scholarly Editing: A Guide to Research.* New York: Modern Language Association of America.

Hughes, L. M. (Ed.). (2012). The Value, Use and Impact of Digital Collections. Facet Publishing.

Kuhn, V. Multimodal. *Digital Pedagogy in the Humanities: Concepts, Models, and Experiments*. https://digitalpedagogy.commons.mla.org/keywords/multimodal/

Lavagnino, J. (2009). Access. Computing the edition. *Literary and Linguistic Computing*, **24**(1): 63–76.

Pierazzo, E. (2015). *Digital Scholarly Editing. Theories, Models and Methods.* Ashgate.

Tanner, S. and Deegan, M. (2003). Exploring Charging Models for Digital Cultural Heritage in Europe. *D-Lib Magazine*, **9**(5). http://www.dlib.org/dlib/may03/tanner/05tanner.html

Vanhoutte, E.(2012). Being Practical. Electronic editions of Flemish literary texts in an international perspective, International Workshop on Electronic Editing (9-11 February 2012), School of Cultural Texts and Records, Jadavpur University, Kolkata, India. http://edwardvanhoutte.blogspot.be/2012/02/being-practical-electronic-editions-of.html

Corpus of Ioannes Dantiscus' Texts and Correspondence dantiscus.al.uw.edu.pl

Anna Skolimowska
annaskolimowska@gmail.com
University of Warsaw, Poland

Ioannes Dantiscus (1485–1548) was a humanist, author of neo-Latin poems, outstanding diplomat in the service of Polish King Sigismund I Jagiellon and Queen Bona Sforza, and then the Senator of the Polish Kingdom as the Bishop of Kulm (1530/33–1537) and Ermland (1537–1548). The final office was also connected with the post of president of the Royal Prussian Council.

Dantiscus' correspondence forms the largest collection of letters (over 6,000) in Central-Eastern Europe, related to the Polish royal court and its partners across the world during that period. Among Dantiscus' 656 correspondents we find rulers, politicians, knights, banquers, humanists and scholars, including especially Nicolaus Copernicus and Erasmus of Rotterdam.

The project "Registration and Publication of the Correspondence of Ioannes Dantiscus", carried out over nearly thirty years at the University of Warsaw (Laboratory for Editing Sources, Faculty of "Artes Liberales"), currently includes two elements:

1. Traditional print publication of the series **Corpus Epistularum Ioannis Dantisci** (appearing in print since 2004 in editorial collaboration between the University of Warsaw and the Polish Academy of Arts and Sciences in Cracow; seven volumes have appeared to date, including three volumes of source texts, two volumes of inventories, and two volumes of studies and commentaries.)

2. Web publication entitled **Corpus of Ioannes Dantiscus Texts and Correspondence** (letters, poems, diplomatic memorials, envoy's speeches and records), dantiscus.al.uw.edu.pl, first published in July 1, 2010.

The "Corpus of Ioannes Dantiscus Texts and Correspondence", with an interface in English and Polish, presents in extenso transcriptions of the primary sources enriched with elaborate critical apparatus. It also contains detailed inventory data on the entirety of Dantiscus' correspondence (sender, addressee, incipit, dating, data on the original source, data on publication in print) and all the Latin texts written by Dantiscus (778 letters, 107 poems, 7 speeches, 15 memorials, 36 records, 1 other text) collected in the "Corpus of Ioannes Dantiscus' Latin Texts", as well as a selection of other correspondence texts (letters to Dantiscus, Dantiscus' letters in German). The selection of texts edited so far was based on the research needs of the program as a whole. Plans for the future envisage publication of the complete "Corpus of Ioannes Dantiscus' German Texts" (letters and documents). The web publication also serves as a first draft for the print publication of the Corpus Epistularum Ioannis Dantisci.

In order to present the source texts online, a custom-designed system of digital registration and annotation for Renaissance correspondence was created, combining transcriptions encoded in accordance with TEI Guidelines, and a relational database that stores a rich set of metadata. All this meticulously collected data offers great potential for further research along many diverse lines. As an example, encoding of references to places and individuals mentioned in this large body of texts allows us to map areas of interest and influence for correspondents or to track their itineraries. Focusing only on people, it is possible to reconstruct parts of social networks of the time, analyzing not only obvious links between correspondents but also references to people co-occurring across the documents. Other areas of scientific investigation have been opened thanks to encoding of variance across extant sources, regularized and original orthography, and other transcriptional features.

The adopted TEI-based approach allows for single source publishing workflow for both the printed series and the web resource, helping to achieve high standards of quality and consistency across publications. As the project and its predecessors span over two decades, technological choices have to be revisited from time to time. The current relational-database-based publishing system, with roots in the early nineties, is planned to be replaced with a native XML database and publishing solution, better suited to satisfying needs for flexible querying of the existing data. Future plans also encompass the integration of visualizations, both to assist research carried out by project collaborators and to provide different ways of engagement with this rich resource, thus broadening its potential audience.

Linking Graph with Map for the Purpose of Historical Research

Jan Škvrňák
jan.skvrnak@gmail.com
Masaryk University, Czech Republic

Adam Mertel
mertel.adam@gmail.com
Masaryk University, Czech Republic

Historical data exploration tool - the middle ages moravian nobility case study

Studies over connectedness of acquired data and their dependence on space and time offer us deeper knowledge of the historical period and also new possibilities of research. In the classical historical science, scholars rarely explore the area between geography (in particular the dependency of phenomenons and people on the environment) and history (social and genealogical context). Thus, we want to create a tool, which allows a parallel observation of geographical and social connections. For this purpose, we will use an experimental data sample of the Moravian nobility in the high and late middle ages. Studies of social elites and their land property offer good conditions for merging several approaches. Simultaneously, there are some difficulties based primarily on the nature of the dataset.

The key part for the creation of a reliable model is the quality of temporal and spatial aspects of the historical data on input. In the 13th century, predicates (nicknames based on the nobility's place of residence) are spreading through the Bohemian lands . Since then we have information about their domains, castles, and strongholds. Further parts of domain as villages were recorded very rarely, because nobility does not need physical documents of their property ownership or property transactions. In contrast, donations from the ruler to the nobility are in better shape. The ecclesiastical institutions have the most detailed evidence of the property because individual monasteries regularly recorded all of their ownership. It was legal protection against a possibility of usurpation by the nobility. The Church evidence can help us with the reverse reconstruction of some nobility estates - if a nobleman had donated or established a monastery or a church. Simultaneously to these phenomenons, the numbers of noblemen began rising and their properties allow us their inclusion into the wider social linkage (as a lineage).

Qualitative change, in field of usable data, brought year 1348, when Land tables (and registry for real estates) was reformed. Two registers for each part of the territory (with capitals in Brno and Olomouc) were established. So called "desky trhové" (libri contractuum, quaterni contractuum) theoretically capture all of the transactions of hereditary estates. Suddenly, we can explore the world of minor trades and speculations of the high and lower nobility, which primarily had not been selling whole domains or at least villages to the Church, but, on the contrary, they had been selling fields or part of villages and their accessories - courts, mills, taverns, and ponds. Next series of the Land book called "knihy půhonné" (libri citationum, quaterni citationum) show us accusations between nobility and in many cases also decisions of the Land Court. Most of the entries is related to a controversial property or unlawful occupancy of villages or their parts, which shows us how much uncertain the property law was. However, accusations about bandit behavior (siege of castles and strongholds, conquering villages, robbery and capturing of subjects) are very rare. Number of social links compared to previous era grew exponentially and for the first time we can observe the nobility as a quarrelling community.

All these specificities of mentioned historical data brought us to an idea of developing an exploratory application that would provide all necessary methods to get a new insight to these datasets. The application itself is based on the combination of modern web technologies (html5 canvas, javascript, d3, leaflet…), visualization techniques and statistical methods. This could bring an complex overview of historical datasets and provide an ideal way to study "hidden" relations. The spatial distribution of data is best explored with the map. For exploration of temporal aspect, the map has to be extended with some cartographic methods and visualization techniques: animation, brushing, filtering or highlighting. These techniques could also be used in the graph view of this application as graph is the most effective way to display and analyse relations over our dataset. In our prototype we will also provide a timeline view which will display the temporal aspect only. The main advantage of this timeline is the possibility to show both duration and existence of phenomenon in particular dates.

This application could be used for further relevant historical research but also as a demonstration of the potential of linking different visualization techniques for extending the functions of historical data exploration. Presented tool will be available on hde.geogr.muni.cz.

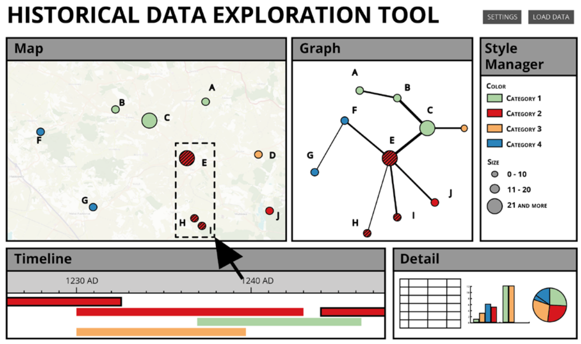


Fig1 : The scheme of the prototype application

Bibliography

Andrienko, N. and Andrienko G. (2006). *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach.* Berlin ; New York: Springer.

Chen, C., Härdle, W. K. and Unwin, A. (2007). *Handbook of data visualization. Springer Science & Business Media.*

Jullien, E. (2013). Netzwerkanalyse in der Mediävistik. Probleme und Perspektiven im Umgang mit mittelalterlichen Quellen. *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte,* **100**(2): 135–53.

Von Lünen, A., Travis, C. (eds). (2013). *History and GIS: Epistemologies, Considerations and Reflection.* Springer, Dordrecht.

Turchin, P., Grinin, L., de Munck, V., Korotayev, A. (eds). (2006). History & Mathematics. *Historical Dynamics and Development of Complex Societies.* KomKniga, Moscow.

New Facets Of The Multimedia Annotation Tool ELAN

Han Sloetjes
han.sloetjes@mpi.nl
Max Planck Institute for Psycholinguistics, Netherlands, The

Olaf Seibert
Olaf.Seibert@mpi.nl
Max Planck Institute for Psycholinguistics, Netherlands, The

Introduction

ELAN is a multimedia annotation tool that is being developed by “The Language Archive” (https://tla.mpi.nl), a department of the Max Planck Institute for Psycholinguistics. It is applied in a variety of research areas within the humanities and beyond; it can be useful in any type of research that includes audio and/or video recordings and analyzes these qualitatively or quantitatively (or both). A general comparison of characteristics of this and other, similar tools can be found in the report of a workshop organized at a gesture conference in Lyon (Rohlfing et al., 2006). This poster with demo is intended as a general introduction to its main functionalities, with an emphasis on the latest developments. Most of these new developments have been executed within CLARIN (Common Language Resources and Technology Infrastructure, http://clarin.eu) projects in the Netherlands and in Germany.

Adding and sharing comments

One of the new developments concerns a commentary framework that improves collaboration of annotators working in a team setting. Comments are pieces of text linked to a segment of the media and possibly to a tier (a tier is an annotation layer). They can be used to store notes, remarks or questions concerning a fragment for later use or for discussion with a colleague. Much alike the way comments in word processors are used. The comments can be shared via email, a file sharing (cloud) service (such as Dropbox) and/or via the back-end of the DASISH Web Annotator (DWAN, http://dasish.eu), a web service for storing annotations (comments) to online content (e.g. web pages). Comments can sometimes be annotations on annotations, but their content and purpose are usually not obvious parts of a (final version of a) transcription.

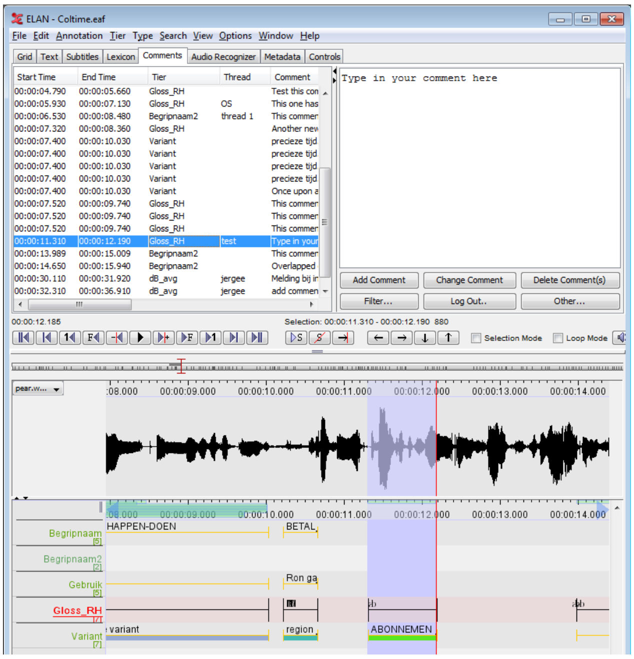
Another recent development is the possibility to associate parts of a transcription to a language identifier (e.g. ISO 639-1/3 code, http://www.iso.org/iso/home/standards/language_codes.htm) in an explicit way. Among these parts are tiers, entries in a controlled vocabulary and individual annotations. The language attribute of tiers can be used for selecting or sorting tiers in the user interface, when exporting or searching.

Connecting to web services

A preliminary implementation of interaction with WebLicht web services (Hinrichs et al., 2010) was presented at the Digital Humanities conference in 2013. Since then this implementation has been modified and updated in several ways. The address of the services called by ELAN are no longer hard wired but instead the user can select a service (representing a parser or tagger etc.) from a list that is obtained from the WebLicht framework itself. The features mentioned above more tightly embed ELAN in the CLARIN world.

Interrater agreement

Other important changes are the new functions for assessing interrater reliability. For many annotation tasks it is important to have an idea of how well annotators are instructed, to what extent annotators agree in their observations and how consistent these are. A simple comparison method solely based on extent and overlap of co-occurring annotations on tiers of two annotators is now complemented by two third party algorithms that take chance agreement into account. One calculates a Cohen's kappa value by first applying a matching algorithm to the segmentations created by two raters (Holle and Rein, 2015). The other calculates a degree of organization by applying Monte Carlo Simulations to segmentations produced by multiple raters (Lücking et al., 2011). It is now possible to perform agreement calculations for an entire corpus, where the user can specify which tiers (i.e. which types of events) need to be assessed.



Start Time	End Time	Tier	Thread	Comment
00:00:04.790	00:00:05.660	Gloss_RH		Test this con
00:00:05.930	00:00:07.130	Gloss_RH	OS	This one has
00:00:06.530	00:00:08.480	Begripnaam2	thread 1	This commen
00:00:07.320	00:00:08.360	Gloss_RH		Another new
00:00:07.400	00:00:10.030	Variant		precieze tijd
00:00:07.400	00:00:10.030	Variant		precieze tijd
00:00:07.400	00:00:10.030	Variant		precieze tijd
00:00:07.400	00:00:10.030	Variant		precieze tijd
00:00:07.400	00:00:10.030	Variant		Once upon a
00:00:07.520	00:00:09.740	Gloss_RH		This commen
00:00:07.520	00:00:09.740	Gloss_RH		This commen
00:00:07.520	00:00:09.740	Gloss_RH		This commen
00:00:11.310	00:00:12.190	Gloss_RH	test	Type in your
00:00:13.989	00:00:15.009	Begripnaam2		This commen
00:00:14.650	00:00:15.940	Begripnaam2		Overlapped
00:00:30.110	00:00:31.920	dB_avg	jergee	Melding bij in
00:00:32.310	00:00:36.910	dB_avg	jergee	add commen

Type in your comment here

00:00:12.185

Selection: 00:00:11.310 - 00:00:12.190 880

A screenshot of the Comments tab and indications of comments in the timeline.

Automatic segmentation and labelling

ELAN allows to create annotations manually, which means that the user can inspect the media stream, identify relevant segments and create annotations using the mouse and/or the keyboard. We have been involved in several projects that aim at integration of tools for semi-automatic segmentation and labelling of the media. A first version was presented in 2013; within the AUVIS (https://tla.mpi.nl/projects_info/auvis/) project the algorithms have been improved and the user interface further streamlined. Information technology experts specialized in analyses of digital video streams closely collaborated with gesture researchers to improve the algorithms for automatic gesture detection and categorization (Schreer et al., 2014) while specialist in speech recognition cooperated with language documentation scientists on better algorithms for speech segmentation and speaker diarization (Rieber and Bardeli, 2013). Although the automatically created segmentation is often not accurate enough to completely replace manually created annotations, these technologies can already be applied in a scenario in which the automatic approach creates the segmentation which is then manually corrected.

Bibliography

Hinrichs, M., Zastrow, T. and Hinrichs, E. (2010). WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*. Valletta, Malta, pp. 489-93.

Holle, H. and Rein, R. (2014). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, **47**(3): 837-47.

Lücking, A., Ptock, S. and Bergmann, K. (2011). Staccato: Segmentation Agreement Calculator. In *Proceedings of the 9th International Gesture Workshop, May 25-27, 2011*. Athens, Greece, pp. 50-53.

Rieber, J. and Bardeli, R. (2013). Speech Recognition as a Retrieval Problem. *Lecture Notes in Informatics*, **220**: 2958-71.

Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., Milde, J.-T., Parrill, F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A. and Wellinghof, S. (2006). Comparison of multimodal annotation tools - workshop report. *Gesprächforschung - Online-Zeitschrift zur Verbalen Interaktion*, **7**: 99-123.

Schreer, O., Masneri, S., Lausberg, H. and Skomroch, H. (2014). Coding Hand Movement Behavior and Gesture with NEUROGES Supported by Automatic Video Analysis. In *Proceedings of Measuring Behavior 2014*, August 27-19, Wageningen, The Netherlands.

Research Data in the Humanities: Present, Preserve and Share

Sibylle Söring
soering@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Mirjam Blümm
bluemm@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Digital resources in the humanities – text, images, audio and video files – show large variation in terms of quality and provenance, while at the same time constantly increasing in number. As a consequence, requirements for digital (and collaborative) editing, annotating, and modelling of data are subject to constant change. Digital editions, for instance, no longer just offer the potential to annotate text and images through established standards (XML/TEI), but also to retrieve entities in an interoperable format (RDF) and to link them to other digital objects (LOD) in other editions, databases, and archives. The digital annotation of both text and/or image data can be published as a digital object itself. Thus, the relation between digital objects turns into a research object in itself, which can be explored, analysed, and searched by a given set of tools and technologies.

A fundamental challenge, however, is that a vast – and still continuously growing – amount of digitised sources is not searchable via one interface or entry point. Search results of numerous different archives, libraries, or other databases cannot be compared automatically; and search requests – as they have to be performed in a de-centralised manner – cannot be (centrally) saved and re-used. Established generic models, together with the necessary expertise in dealing with individual project data as digital objects, are still lacking practice and broad usage in digital humanities and cultural studies. This applies for both data modelling as well as for long-term archiving and persistent publishing of primary digital data.

The proposed poster will illustrate these core aspects of generating, modelling, archiving, exploring, and presenting humanistic research data from a structural and methodological perspective. The digital infrastructure DARIAH-DE is developing a repository for humanities' research data (https://de.dariah.eu/repository) to be archived and published in a sustainable and persistent way, allowing for re-use, citation, preservation, and further exploration:

- Persistent Identifiers, e.g. Digital Object Identifiers (DOI), enable scholarly users to identify research data distinctly. They serve as a crucial pre-condition for archiving, retrieving and citing primary digital data not only in science, but in the humanities as well.
- The modelling of digital objects with established standards (CIDOC, SKOS etc.) and formats (RDF) provides the basis for generating links, thus allowing for generating new knowledge on objects, collections, and/or corpora. Modelling data with the RDF standard allows for linking different objects and therefore accessing information, which can be encoded in variable formats and stored in different databases and archives. The poster will show some practical examples for generic search across TEI-encoded corpora.
- Digital data collections can serve as innovative sources for humanities' research. As such, they can potentially generate new knowledge. Which are mandatory prerequisites for dealing adequately with information in image- and object- orientated research? Which standards for structuring research data apply, enabling a comprehensive search functionality also across heterogeneously structured data collections and archives? The poster will also present the DARIAH-DE Data Federation Architecture, a set of different tools and services creating innovative options of data management, e.g. accessing and linking heterogeneous data sources of various provenances, analysing existing distributed data collections as well as services to generate data and schema interoperability.
- The poster will show how repositories can help users of varying research disciplines to model, store, and publish their data in a standardised, interoperable way, allowing for re-use, sharing, and further exploration.

The poster will accompany a hands-on workshop adressing these questions in more detail, together with (potential) users. DARIAH-DE will present its repository, its features, the Data Federation Architecture and other tools and services related to dealing with and modelling of humantistic research data via demonstrations and hands-on sessions, enabling participants to actively explore the services also with their own data.

The Latin Batrachomyomachia Collection

Petra Sostaric
petrasostaric2011@gmail.com
University of Zagreb, Croatia

Sinisa Jovcic
sinisa.jovcic74@gmail.com
VERN University of Applied Sciences, Croatia

The topic of our project follows closely the theme of this conference: the digital humanities have been a bridge between the past and the future in the field of classical philology, and the identity of the classical scholar has imperceptibly but steadily been merging with a digital humanist identity. As humanist translations from Greek to Latin receive more and more attention, it would be useful to have translations of specific texts collected, organized and made machine-searchable to enable further scholarly study, instead of having them scattered as scanned manuscripts in different repositories. Usually text collections offer a corpus of national literature (e.g. *CroALa*) or specific periods (e.g. *Biblioteca Digital del Pensamiento Novohispano*). This project focuses on a single text: the aim of this project is to produce an online collection of Latin translations of the Greek mock-epic *Batrachomyomachia*, consisting of 300 lines. The *Batrachomyomachia* had long been attributed to Homer and attracted much attention as such; it was even considered Homer's best work by some. The revival of Greek studies in the West in the Renaissance lead to a lively production of translations from ancient Greek into Latin, even of what is now considered minor works and authors. The interest did not disappear after the Renaissance and the production of humanist translations of different works continued. The last Latin translation of the *Batrachomyomachia* was to appear in the 18th century, penned by a renowned Mexican humanist, Francisco Javier Alegre. Unfortunately, humanist translations are seldom included into existing online text collections. The database of Italian Renaissance Latin poetry, *Poeti d'Italia*, does not include a Latin *Batrachomyomachia* under the name of Renaissance humanist Carlo Marsuppini, nor is there the Latin *Batrachomyomachia* by Joachim Münsinger von Frundeck at the *CAMENA* database of neo-Latin poetry. The interest of modern day digital humanists for this mini-epic does not reciprocate the interest of the earlier humanists who used Latin as their medium of choice, but as translations from Greek to Latin have been coming into classicists' focus in recent years, this kind of collection will certainly prove itself useful.

The aim of our project is to collect as many Latin translations of the *Batrachomyomachia* as possible (many are already in the public domain as they have been digitised by various institutions; they will be transcribed by researchers in this project) and make them machine searchable and available online for further scholarly study. Our application will be used for uploading .docx documents and their conversion into xml files while erasing unnecessary tags. Every word will be saved in a lexicon and a table with information on the word's position in the lexicon and in the text. The search will be conducted through a MySql query with word and context as a result. The project will also include Greek to Latin text alignments (micropublications done in *Perseids*) and educational material: quizzes developed especially for teaching Greek while enhancing the learner's Latin skills with material taken from the Greek and Latin *Batrachomyomachias*. The material collected in this project will be useful to teachers, students and researchers.

Bibliography

Heidelberg University 2013. *CAMENA*. Available at: http://www.uni-mannheim.de/mateo/camenahtdocs/camena_e.html. Accessed 6. 3. 2016.

Tufts University 2007. *Perseus Digital Library: Perseids*. Available at http://perseids.org/. Accessed 6. 3. 2016.

UNAM 2016. Biblioteca Digital del Pensamiento Novohispano. Available at: http://www.bdpn.unam.mx/. Accessed 6. 3. 2016.

University of Zagreb 2014. CroALa. Available at: http://croala.ffzg.unizg.hr/. Accessed 5. 3. 2016.

University Ca' Foscari 2016. *Poeti d'Italia*. Available at: http://www.poetiditalia.it/public/. Accessed 5. 3. 2016.

Complex Networks-Based Approach to Russian Rhyme History Description: Linguostatistics and Database

Olga Sozinova
oa.sozinova@gmail.com
Higher School of Economics, Moscow, Russian Federation

Introduction

Russian rhyme was described thoroughly in the 20th century, especially by M. Gasparov (Gasparov, 2000). Though we now have a powerful tool to analyze rhymes further in the form of the poetic corpus of the Russian National Corpus (henceforth RNC), not much recent research has been carried out in this area (Orekhov, 2015). As I am particularly interested in visualizing corpus data, I applied graphs to rhyme analysis.

Rhymes are convenient entities to be described in graph terms. In a rhyming pair, words are nodes, and rhyme relationship between them is an edge between nodes. Certain properties can be assigned to the nodes and to the edges. For example, word nodes may contain grammatical information and rhyme edges may bear all the rhyme classification (meter, position, etc.).

Furthermore, nowadays, there are tools available for storing graph information in a database. Information from such databases can be retrieved easily and in several formats.

My aim was to build a graph database using the data from the poetic corpus of the RNC. I want to show that the manual research done previously can be supported and extended in a vivid graphic way.

Graphs can provide us with much information about rhyme diachrony:

- Degree of connectivity in different periods (different rhyming tendencies);
- The longest path (chain of rhymes) and clusterization (popular rhymes in different periods);
- Tendency flow from exact rhymes to inexact (requested by parameter of exactness in the different periods);
- Appearance of the dissonance rhymes;
- Tendency flow in rhyming types, position;
- Domination of a certain rhyming type within the rhymes of one poet.

Data

The whole poetic corpus of the RNC was used for analysis. The data covers 775 Russian authors, born between 1658—1939. Overall, the corpus contains 85,996 documents, 229,968,798 words.

Analysis

Technical work included the following steps:

- Transcribing words in a rhyme position;
- Retrieving rhymes from poems according to the phonetic transcription;
- Rhyme classification;
- Building new nodes and edges in the graph database.

I used Python for rhyme extraction and classification and a Neo4j database for storing the data.

As I could not find any available modules for Russian transcription, I made this module myself using the transcribing rules from [http://www.philol.msu.ru/~fonetica/index.htm]. The module takes into account almost every rule, but the exception word lists are quite small.

The rhyme extraction algorithm I used was the following. The Python program finds all the tags <rhyme-zone> (tagged everywhere except blank verse) in the XML documents with poems (last words in lines). Then the program tries to find a possible rhyming pair within 4 lines before and after the current rhyme-zone word. Afterwards comes the transcriptions comparison; if stressed vowels are the same, then the process of classification begins. If the stressed vowels are different, then the dissonance rhyming type is checked.

The classification of the rhymes was based on (Surkov, 1962), (Kvjatkovskij, 1966) and (Timofeev, 1935). I took 8 parameters into account:

- Exactness (exact or inexact);
- Richness (rich or poor);
- Depth (deep or not);
- Ending (open or closed);
- Position of the stressed vowel (male, female, dactyl, hyperdactyl);
- Rhyming type (paired, crossed, encircling);
- Assonance;
- Dissonance.

As soon as a new rhyming pair is found and classified, new nodes and edges are automatically created in the graph database. If any of the words existed in the database before, then an edge is created to the existing node; otherwise, a new node is created.

Results

For now, I have rendered several graph images for certain poets with approximately 30% of their rhymes. In Figure 1 there are 2570 rhymes from the poems of P. Vjazemskij. Figure 2 shows 3866 rhymes from the poetry of A. Pushkin.

From the figures, we can see that connectivity in Pushkin's graph is much higher than in Vjazemskij's graph. Furthermore, the graph of Vjazemskij's rhymes demonstrates certain clusters which can be analyzed in detail.

I plan to continue my research and obtain other graphs for the whole epoch. I hope that further work will provide the information I listed in the introductory part, especially regarding connectivity and clusterization over different time periods. Quantitative analysis remains to be done as well. Firstly, I would like to look at the graph patterns, and then go deeper into calculations of graph characteristics and their interpretations.

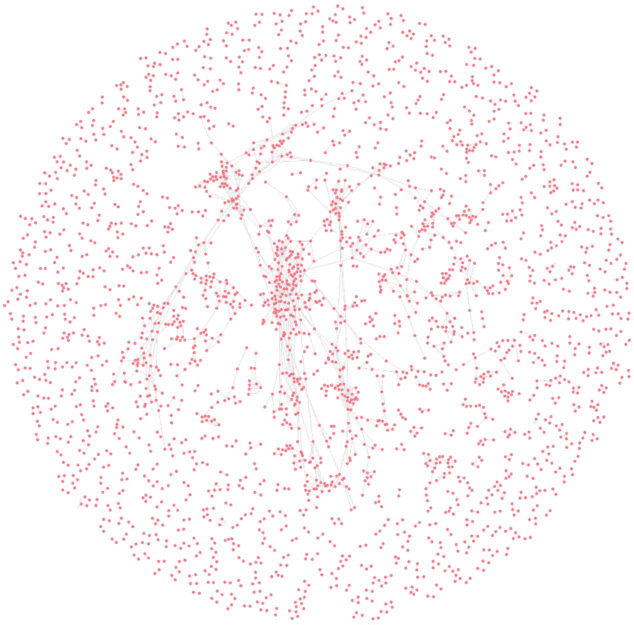


Figure 1. Graph of the 2570 P. Vjazemskij's rhymes

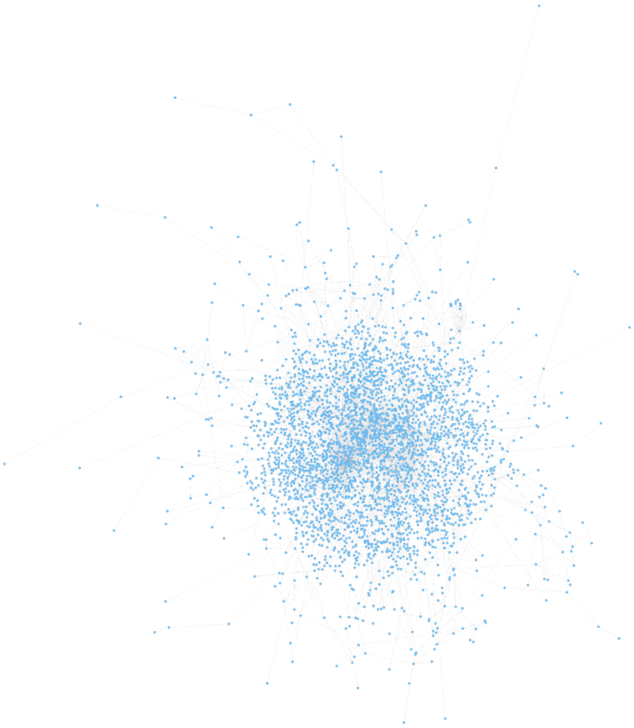


Figure 2. Graph of the 3866 A. Pushkin's rhymes

Bibliography

Gasparov, M. (2000). *Očerk istorii russkogo stikha* [Studies of the Russian verse history]. Fortuna limited.

Kvjatkovskij, A. (1966). *Poetičeskij slovar'* [Poetic dictionary]. Sovetskaja Enciklopedija.

Orekhov, B. (2015). *Ešče raz ob issledovatel'skom potenciale poetičeskogo korpusa: metr, leksika, formula* [One more time about the research potential of the poetic corpus: meter, lexicon, formula]. Russian National Corpus, in print.

Surkov, A. (1962). *Kratkaja literaturnaja enciklopedija* [Short literary encyclopedia], 1. Sovetskaja enciklopedija.

Timofeev, L. (1935). *Literaturnaja enciklopedija* [Literary encyclopedia], 9. Sovetskaja enciklopedija.

Cantus Network – a Semantically Enriched Digital Edition of Libri Ordinarii of the Salzburg Metropolitan Province

Christian Steiner
christian.steiner@uni-graz.at
Karl-Franzens-Universität Graz, Austria

Introduction

For many centuries, the metropolitan province of Salzburg played a key role in the cultural history of Austria. The digital availability of the many surviving liturgical musical sources which form an important part of this cultural heritage is therefore of great importance. This research project aims at investigating the records that survived as manuscripts and describe the practice of liturgical and musical acts of worship. Liturgical ordinals, called libri ordinarii (LOi), are key sources for this transmission, as they include a short form of the entire rite of a diocese or a monastery. Prayers, readings and chants are given as abbreviations (incipits).

This poster examines the technical background to the project and specifically the transformation process into TEI as well as the creation of semantic web knowledge representations of the liturgical occasions/functions and the chants based on existing relational databases.

The object: Liturgy and music in the medieval metropolitan province of Salzburg

A liber ordinarius usually includes all the information required by an individual institution (church, monastery) or a group (diocese, group of monasteries) for their services. This includes the incipits of chants, readings and prayers for the liturgy of the hours, mass and processions as well as rubrics that provide instructions on how and when particular liturgical acts should be carried out.

The rubrics often contain indications of how the chant is to be performed and are thus able to provide important information on chant performance practice. However, this information cannot be utilized without in-depth examination of the LOi: thus, scholarly studies have to include a critical transcription of the Latin texts, followed by an in-depth analysis of the origins of the liturgy and the commentaries.

The aim: Reconstruct the development of the liturgy in the Salzburg metropolitan province by Semantic Enrichment

The research project's aim is to reconstruct and produce a synoptic study of the emergence and development of the liturgy in the Salzburg metropolitan province, based on the surviving LOi in the region. A primary task was the transformation of the ordinarii from docx into TEI. A Java algorithm parses the docx on word-level in order to create the desired TEI-Output. The encoding in TEI consequently allows a detailed transcription of textual phenomena and the reference between texts. This is the basis for an in-depth analysis of the different traditions and variants of the LOi.

The LOi as a textual genre need enriched editions, which cannot be fully provided in printed forms. The project will provide the multi-layered texts digitally working with a controlled vocabulary. This resource will be of particular interest for the international efforts on cataloguing the manuscript heritage and research on the history of liturgy, which will be able to refer to the controlled vocabulary via Semantic Web technologies and Linked Open Data. The data from the existing relational databases cantusplanus.at and cantusdatabase.org will be converted into semantic web knowledge representations (SKOS, OWL). With the usage of these standards, we can secure the long-term use of the data created.

The long term perspective: The technical infrastructure

The digitized and enriched objects will be managed, published and long-time archived in GAMS (Geisteswissenschaftliches Asset Management System). GAMS is an OAIS-compliant Asset Management System based on the Open Source software FEDORA and further developed by the Austrian Centre for Digital Humanities at the University of Graz.

GAMS focuses on the long-term availability and flexible use of digital content. The repository builds upon a webservice-based (SOAP, REST), platform-independent and distributed system architecture, a largely XML based content strategy, the support of XML based import and export standards (METS, DIDL, etc.) and the use of standardized data and metadata formats. All data objects in the system receive a persistent identifier (PID) and can thus be explicitly cited.

During the ingest process of the objects into the repository, the TEI documents can be semantically enriched through content-specific controlled vocabularies. Information will be extracted from the content document, for instance Dublin Core metadata or user-generated RDF triples for subsequent processing like complex search requests, the generation of indices and data visualization. For the storage and retrieval of RDF triples, an openRDF Sesame repository is accessed through a web service.

The manuscripts in the project will be presented with an interface, which allows navigating the data in the ways mentioned above, including elaborate multi-criteria search support by drill down methods. GAMS includes an efficient Imageserver (IIPImage) which allows seamless navigation and zoom in high resolution images.

Bibliography

Klugseder, R. (2014). Mittelalterliche musik-liturgische Quellen aus dem Augustinerchorherrenstift St. Florian. In *Musicologica Austriaca*, **31**.

Praßl, F. K. (1998). Der älteste Salzburger Liber Ordinarius (Codex M II 6 der Universitätsbibliothek Salzburg). Zu seinen Inhalten und seiner Wirkungsgeschichte. In Engels, S. and Walterkirchen, G. (eds.), *Musica sacra mediaevalis. Geistliche Musik Salzburgs im Mittelalter*. St. Ottilien: pp. 31-47 (= Studien und Mitteilungen zur Geschichte des Benediktinerordens und seiner Zweige 40, Ergänzungsband)

CANTUS. (2016). A database for latin ecclesiastical chant. http://cantusdatabase.org/ (accessed 22 February 2016)

GAMS. (2016). http://gams.uni-graz.at/doku (accessed 22 February 2016)

EDM in Use: Collecting Metadata for a Regional Cultural Heritage Portal

Elisabeth Steiner
elisabeth.steiner@uni-graz.at
University of Graz, Austria

The Europeana Data Model (EDM) aims to provide an abstract and formal specification for the delivery of data to Europeana. The model is meant to replace the older Europeana Semantic Elements (ESE) definition. Though Europeana still accepts data provided in ESE, the newer model means to provide a richer set of description possibilities and fine granulation in the distinction of the provided digital heritage object (provided CHO) and one or more digital representations thereof. Most notably, EDM is designed to provide Linked Open Data (LOD) of the described resources (cf. Berners-Lee, 2006). Thus, data providers are encouraged to switch to EDM and to take advantage of the extended potential.

The project "Repository of Styrian Cultural Heritage" will build a digital archive of cultural heritage objects, i.e. a common web platform where all partners (3 universities, 2 museums, and the local government) will expose metadata on their Styrian specific collections and holdings. As the nature of digitized objects ranges from text-centered materials like manuscripts or correspondence to images like historical photographs to museum objects of various contexts, an abstract data model for representation and retrieval was needed.

In a first step, the metadata core categories obligatory for all partners and content types were defined: They roughly correspond to Dublin Core (DC) and DCTerms and will form the basis of the web portal: institution, title/description, person/creator, time, place, object type, media type.

In a second step, this abstract model had to be converted into a formal specification of how to provide the metadata for a harvesting mechanism. The process of choice was OAI-PMH, either in form of an OAI-PMH interface or as an OAI 2.0 compliant XML-file. Harvesting of metadata on diverse cultural heritage objects naturally calls for the application of EDM expressed in XML. Thus, the abstract metadata core categories are mapped to EDM properties and formalized as XML elements and attributes. EDM is a powerful yet flexible tool, therefore a standardized application profile had to be developed for the project.

An important consideration is the integration of controlled vocabularies and norm data. EDM offers the possibility to do so, yet the incorporation of geographical coordinates for instance calls for the modelling of the place name in the edm:Place element and not in dc:coverage or dcterms:spatial.

At the end, the EDM XML data is integrated into the OAI-PMH stream. The goal is to arrive at a recommendation on data structure for the contributing institutions, a process which is currently underway.

Using EDM offers obvious advantages in the context of a harvesting portal: as it was specifically designed to capture the field of data aggregation from various sources, it offers a good spectrum of possibilities to address the difference between the original (i.e. analog) cultural heritage object (edm:providedCHO) and available web resources (edm:WebResource), but also sets these two aspects in correlation (ore:Aggregation).

The conception of a least common denominator template is achieved relatively easy though this is surely not a trivial task, taking into consideration the flexibility and integration potential of the model. Another important aspect needs constant supervision during the project: maintaining metadata quality and how to fill the template accordingly. What type of annotation of person forenames and surnames should be used? What set of keywords? What date format? What controlled vocabularies are vital? How to deal with uncertainties and fuzziness inevitably occurring in such datasets? These questions need to be addressed in cooperation with the scholars working on the collections, clarifying that these annotations will determine the accurateness of search and the quality of data retrieval and representation on the web portal. With regard to the (re-)use as Linked Open Data, special attention is given to the annotation with suitable norm data.

EDM offers a good framework for use in harvesting contexts, especially where various content types are present. Nevertheless, the experience in the project has shown that offering a technical concept is only half of the story: To achieve formal data interoperability and to simultaneously populate this model with comparable content categories is the great challenge in this context. Only with such a homogenous data basis the resulting web portal will be able to offer features like timelines, map-oriented visualisations and other discovery mechanisms. The poster will introduce the project's so far approaches, solutions and lessons learned from this point of view.

Bibliography

Berners-Lee, T. (2006). Linked Data. https://www.w3.org/DesignIssues/LinkedData (accessed 29 February 2016)

DC and DCTerms. http://dublincore.org/documents/dcmi-terms/ (accessed 29 February 2016)

Definition of the Europeana Data Model v5.2.6 (2014). http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM%20Definition%20v5.2.6_01032015.pdf (accessed 29 February 2016)

EDM Documentation. http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation (2016-02-29)

Europeana.http://www.europeana.eu/portal/ (accessed 29 February 2016)

Europeana Semantic Elements Specification and Guidelines (2013). http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/ESE_Documentation//Europeana%20Semantic%20Elements%20Specification%20and%20Guidelines%2014%20July%202013.pdf (accessed 29 February 2016)

OAI 2.0.http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed 29 February 2016)

OAI-PMH.http://www.openarchives.org/pmh/ (accessed 29 February 2016)

Repository of Styrian Cultural Heritage.http://wissenschaftserbe.uni-graz.at/en/ (accessed 29 February 2016)

Making George Washington's Financial Documents Accessible: Transcription, Data, and the Drupal Solution

Jennifer Elizabeth Stertzer
jes7z@virginia.edu
University of Virginia, United States of America

Erica Fallon Cavanaugh
efc8d@virginia.edu
University of Virginia, United States of America

The George Washington Financial Papers Project exists at the intersection of two challenges editors currently face: managing complicated editorial work and navigating the world of digital publication. By focusing on a particularly difficult and dynamic dataset—financial documents—work has advanced on three interconnected fronts: 1) developing document templates for both traditional financial documents, such as account books and ledgers, as well as receipts, journals, and memoranda; 2) developing taxonomies and data visualizations; and 3) constructing an open-source content management/editorial/publication platform. The work has resulted in the development of both an open-access digital edition of Washington's financial documents as well as the groundwork of Drupal for Editors prototype—a Drupal-based, open-source, editorial/publication platform—providing editors with a stable, flexible, and powerful platform to build engaging digital editions of financial documents.

In 2013, The Papers of George Washington received a grant from the National Historical Publications and Records Commission (NHPRC) for the perfection and population of the content management database (DocTracker) with Washington's three major ledger books; preparation of Gouverneur Morris's 1811-1816 account book and its entry into the content management database in partnership with the Gouverneur Morris Papers at the New-York Historical Society; and the completion of a primary version of a web interface that will provide users with free access to the edition's entire content and permit downloading and data manipulation.

During our partnership with DocTracker we helped design a viable content management and customized editorial workflow solution built on the proprietary, commercial database software FileMaker Pro. DocTracker allowed us to manage both document records and content identifications, and associate both with transcriptions. But as a publication platform it was limited because of its use of XML. We investigated alternative publication options and decided on Drupal, a highly-configurable open-source content management system. We determined Drupal to be the best publication solution for several reasons: 1) at its core, Drupal is a database in which imported content can be mapped to fields, allowing for robust displays and searching, querying, and browsing; 2) Drupal is accessible, both in terms of cost and usability and has a large user community; 3) both the backend (content/data) and frontend (website interface) are managed in the system; and 4) Drupal is open-source and its core and add-on (module) code are developed and actively maintained by a large international developer community.

Drupal has allowed the project to confront the numerous challenges inherent in these documents: (1) different types of financial documents are formatted in distinct, though standardized, ways, and the formatting of financial documents carries implied meanings; (2) transactions are full of dittos, abbreviations, and short hand, that raise a question of what kind of fields should be created to capture the transcription and clear text, thereby making both the text and content searchable; (3) the documents present issues of currency, valuation, and barter; and (4) a hierarchy of documents exist, and therefore the same transaction may be recorded in a day book, account, and ledger, etc., generating multiple instances of the same transaction.

Indeed, one of the primary goals of the Project is to make accurate transcriptions of the documents *available*, in keeping with the long tradition of the Papers of George Washington documentary editing project. However, the types of information, or the "data," contained in these documents are not easily *accessible* using common search and query techniques. The challenges, as described above, make it impossible to simply transcribe and put online, ready to be searched and understood document transcriptions. The solution involves a combination of transcription and corresponding data fields (where dittos, abbreviations, and short hand have been expanded), node references associating various content types, and term references connecting taxonomies. Additionally, Drupal provides a place to develop and manage taxonomy lists for specific content types, such as financial documents, to enhance the grouping and sorting of content and be used to identify relationships between different types of content.

Developing this system has challenged us to think creatively about all aspects of the editorial and publication process, resulting in innovative ways for users to explore, analyze, and interact with content. This poster and hands-on demonstration will explore these issues and the technological solutions to make these documents available, as a free online resource as well as highlight strategies for content searchability, including annotation, glossaries, indexes, and linking.

The Models of Authority Project: Extending the DigiPal Framework for Script and Decoration

Peter Anthony Stokes
peter.stokes@kcl.ac.uk
King's College, London, United Kingdom

Stewart J. Brookes
stewart.brookes@kcl.ac.uk
King's College, London, United Kingdom

Geoffroy Noël
geoffroy.noel@kcl.ac.uk
King's College, London, United Kingdom

John Reuben Davies
John.R.Davies@glasgow.ac.uk
University of Glasgow, United Kingdom

Tessa Webber
mtjw2@cam.ac.uk
University of Cambridge, United Kingdom

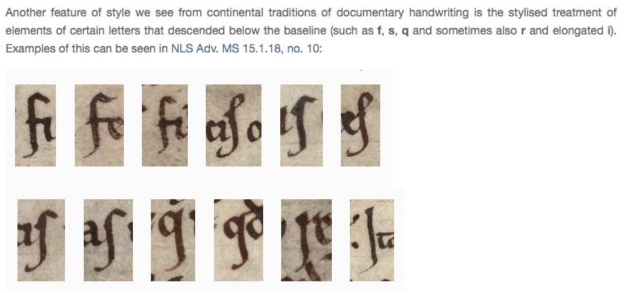
Dauvit Broun
Dauvit.Broun@glasgow.ac.uk
University of Glasgow, United Kingdom

Alice Taylor
alice.taylor@kcl.ac.uk
King's College, London, United Kingdom

Joanna Tucker
Joanna.Tucker@glasgow.ac.uk
University of Glasgow, United Kingdom

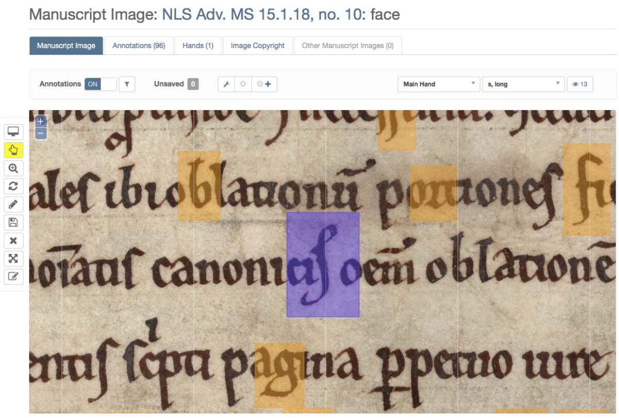
The DigiPal project for palaeography has featured in previous DH conferences (Stokes 2012; Stokes *et al.* 2014; Stokes 2015). It includes a generalised framework for the description and analysis of handwriting, initially applied to Old English of the eleventh century but subsequently extended to Latin, Hebrew, and decoration (Stokes *et al.* 2014); it incorporates a novel model for describing handwriting (Stokes 2012); and a recent addition allows the embedding of linked palaeographical images into prose description (Figures 1 and 2; see also Figure 3). The purpose of this poster is to present new developments which form part of two further major grants, one of which is the Models of Authority project (see also Exon 2015). Specifically, the focus here is on the incorporation of textual content into the model for handwriting.

It is well known that features of handwriting depend not only on the type of manuscript and the social context of production, but also on the text itself, and the absence of this has always been a weakness of DigiPal (Stutzmann 2012; Stutzmann 2013). For Models of Authority, in contrast, a key premise is that the development of government in medieval Scotland is reflected in the choice of visual models behind the layout and handwriting of royal and local documents: this choice probably drew primarily on English counterparts, but elements from papal bulls, among others, are also visible (Broun 2015). It is helpful to note further that the text of these documents is relatively formulaic: this is common of charters in general, as noted in particular by the Charters Encoding Initiative (CEI) and utilised effectively by the Anglo-Saxon Charters project (ASChart), among others. However, connections between diplomatic formulae and script remain largely unexamined, at least in a digital context; certainly digital projects are underway that analyse the relationship between script and document (two examples are PUhMA and MoM-CA), but freely-available online frameworks are still lacking that allow queries like 'show me images of **d** that appear in witness-lists of episcopal charters, and let me compare them with those of dispositive clauses'. Hence a key question of Models of Authority is whether matching the palaeography to the diplomatic formulae in this way might reveal new information about how the documents were put together, what models may lie behind them, and therefore how the nascent Scottish government may have seen itself particularly in relation to the English and Papal chanceries.



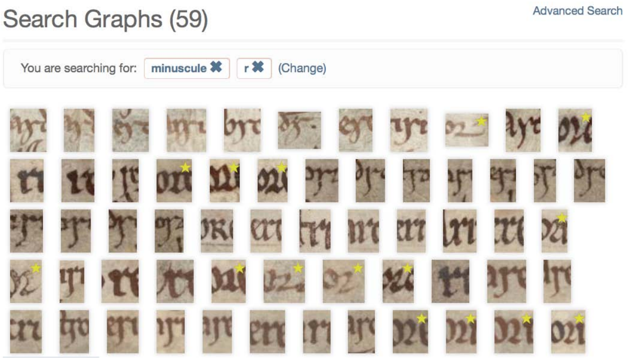
Another feature of style we see from continental traditions of documentary handwriting is the stylised treatment of elements of certain letters that descended below the baseline (such as **f**, **s**, **q** and sometimes also **r** and elongated **i**). Examples of this can be seen in NLS Adv. MS 15.1.18, no. 10:

Figure 1: Sample of Embedded Annotations in Palaeographical Description (Webber 2015). All images are linked by the framework to their full context in the manuscript page.



Manuscript Image: NLS Adv. MS 15.1.18, no. 10: face

Figure 2: The Graph of **s** (fourth across in Fig. 1 above) in its page context.

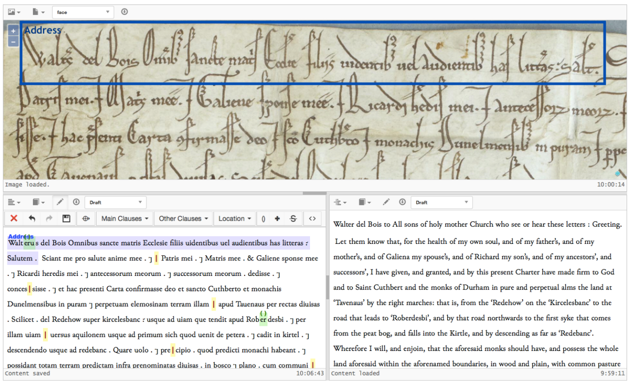


Search Graphs (59)

You are searching for: minuscule ✖ r ✖ (Change)

Figure 3: Sample forms (graphs) of **r** in the Models of Authority database. Examples of 2-shaped **r** are starred.

The DigiPal framework has therefore been extended by adding a generalised textual component, such that arbitrary marked-up texts (transcription, translation, codicological description, and so on) may be associated with palaeographical annotations and images and displayed in parallel with them (Figures 4–6). This allows the exploration of correlations between palaeographical forms and the corresponding text, comparing in this case the different clauses in the documents not just in terms of the text (as done in ASChart and others) but also as images of the corresponding clause in different documents. The integration of this into the DigiPal model for handwriting (described by Stokes 2012) therefore allows detailed palaeographical querying, such as filtering for images of all the letters with ascenders that appear in the address clause of the charter, or looking for variation in letter-forms in the witness lists. The former is of considerable interest because chancery scribes often add decorative elements to ascenders of letters, particularly in the first line of the text (visible in Figure 4), a practice which derives from Papal and ultimately Merovingian diplomatic practice; for the latter query, names of witnesses are normally written by the main scribe at the same time as the rest of the text but in some cases may also have been added later (Broun 2011, 258–65 and 280–7), so identifying differences in writing here can also be very significant. Beyond charters, the digital model generalises to other forms of linking from structured text to structured annotation, for instance at the glyph-by-glyph level advocated by Stutzmann (2013), and is also being applied elsewhere (Exon 2015).



Address
Walterus del Bois Omnibus sancte matris Ecclesie filiis uidentibus uel audientibus has litteras : Salutem . Sciant me pro salute anime mee . ⁊ | Patris mei . ⁊ Matris mee . & Galiene sponse mee . ⁊ Ricardi heredis mei . ⁊ antecessorum meorum . ⁊ successorum meorum . dedisse . ⁊ conces|sisse . ⁊ et hac presenti Carta confirmasse deo et sancto Cuthberto et monachis Dunelmensibus in puram ⁊ perpetuam elemosinam terram illam | apud Tauenaus per rectas diuisas . Scilicet . del Redehow super kircelesbanc : usque ad uiam que tendit apud Roberdesbi . ⁊ per illam uiam | uersus aquilonem usque ad primum sich quod uenit de petera . ⁊ cadit in kirtel . ⁊ descendendo usque ad redebanc . Quare uolo . ⁊ pre|cipio . quod predicti monachi habeant . ⁊ possidant totam terram predictam infra prenominatas diuisas . in bosco ⁊ plano . cum communi |

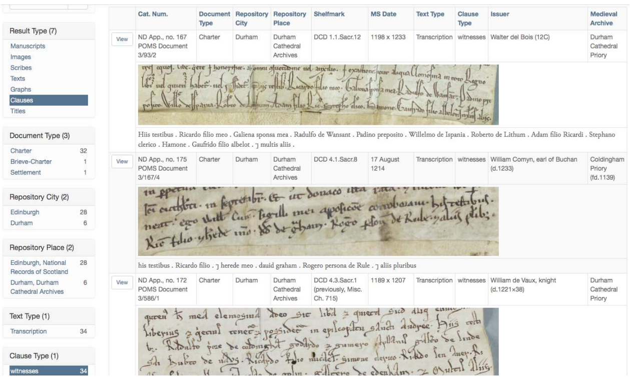
Content saved 10:06:43

Walter del Bois to All sons of holy mother Church who see or hear these letters : Greeting.

Let them know that, for the health of my own soul, and of my father's, and of my mother's, and of Galiena my spouse's, and of Richard my son's, and of my ancestors', and successors', I have given, and granted, and by this present Charter have made firm to God and to Saint Cuthbert and the monks of Durham in pure and perpetual alms the land at 'Tavenaus' by the right marches: that is, from the 'Redehow' on the 'Kircelesbanc' to the road that leads to 'Roberdesbi', and by that road northwards to the first syke that comes from the peat bog, and falls into the Kirtle, and by descending as far as 'Redebanc'. Wherefore I will, and enjoin, that the aforesaid monks should have, and possess the whole land aforesaid within the aforenamed boundaries, in wood and plain, with common pasture

Content loaded 9:59:11

Figure 4: Diplomatic Markup of Text and Image



Result Type (7)
- Manuscripts
- Images
- Scribes
- Texts
- Graphs
- Clauses
- Titles

Document Type (3)

Charter	32
Brieve-Charter	1
Settlement	1

Repository City (2)

Edinburgh	28
Durham	6

Repository Place (2)

Edinburgh, National Records of Scotland	28
Durham, Durham Cathedral Archives	6

Text Type (1)

Transcription	34

Clause Type (1)

witnesses	34

	Cat. Num.	Document Type	Repository City	Repository Place	Shelfmark	MS Date	Text Type	Clause Type	Issuer	Medieval Archive
View	ND App., no. 167 POMS Document 3/93/2	Charter	Durham	Durham Cathedral Archives	DCD 1.1.Sacr.12	1198 x 1233	Transcription	witnesses	Walter del Bois (12C)	Durham Cathedral Priory
	Hiis testibus . Ricardo filio meo . Galiena sponsa mea . Radulfo de Wansant . Padino preposito . Willelmo de Ispania . Roberto de Lithum . Adam filio Ricardi . Stephano clerico . Hamone . Gaufrido filio albelot . ⁊ multis aliis .									
View	ND App., no. 175 POMS Document 3/167/4	Charter	Durham	Durham Cathedral Archives	DCD 4.1.Sacr.8	17 August 1214	Transcription	witnesses	William Comyn, earl of Buchan (d.1233)	Coldingham Priory (fd.1139)
	his testibus . Ricardo filio . ⁊ herede meo . dauid graham . Rogero persona de Rule . ⁊ aliis pluribus									
View	ND App., no. 172 POMS Document 3/586/1	Charter	Durham	Durham Cathedral Archives	DCD 4.3.Sacr.1 (previously, Misc. Ch. 715)	1189 x 1207	Transcription	witnesses	William de Vaux, knight (d.1221×38)	Durham Cathedral Priory

Figure 5: Example query showing text and image of witness lists in private charters.

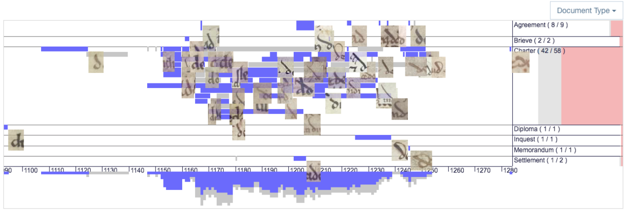


Figure 6: Prototype overview/timeline indicating shapes of **d** in documents of different type.

This addition of text introduces considerable complexity to the system, since it seeks to connect very different views of the original document: as transcription, translation, image and annotation; as scribal product (a set of interconnected and interacting letters) and diplomatic product (a set of textual formulae); an overview of the current architecture is shown in Figure 7. This integration of different views goes to the heart of an ongoing challenge in Digital Humanities, as perhaps best represented by the competing 'textual' and 'documentary' views in the TEI (2015; Pierazzo and Stokes 2011; see also Stokes and Noël 2010).

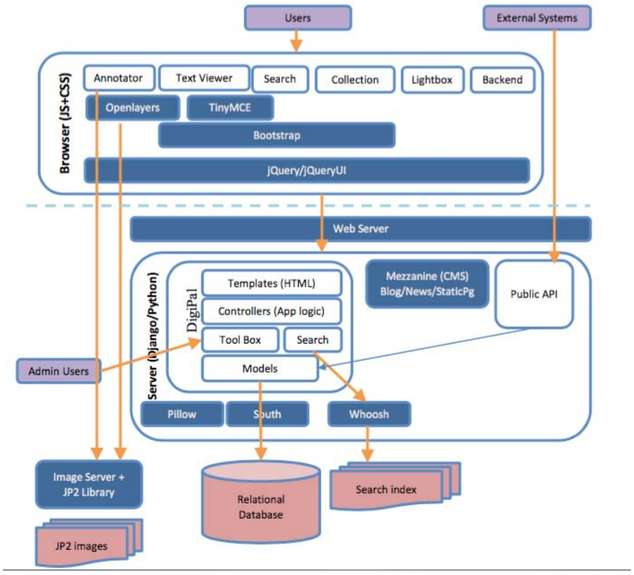


Figure 7: Overview of the current architecture

The proposed poster will therefore present the existing framework and integrated model, addressing the complexities described above. Live demonstrations will be available of Models of Authority, and of other instances of the framework such as DigiPal, Exon, and some that are not yet publicly available.

Acknowledgements

The Models of Authority project is funded by the Arts and Humanities Research Council (AHRC) under Grant Reference n° AH/L008041/1.

Bibliography

ASChart. *Anglo-Saxon Charters*. London: King's College. Available at http://aschart.kcl.ac.uk/. [Accessed 1 November 2015].

Broun, D. (2015). Introducing the 'Models of Authority' project: Scottish charters c. 1100–c.1250. *Models of Authority: Scottish Charters and the Emergence of Government*. London: King's College. Available at http://www.modelsofauthority.ac.uk/blog/intro/. [Accessed 1 November 2015].

Broun, D. (2011). The presence of witnesses and the writing of charters. In Broun, D. (ed.), *The Reality behind Charter Diplomatic in Anglo-Norman Britain*. Glasgow: University of Glasgow, pp. 235–90. Available at http://paradox.poms.ac.uk/redist/pdf/chapter4.pdf. [Accessed 3 March 2016].

CEI. *Charters Encoding Initiative*. Munich: Ludwig-Maximilians Universität. Available at http://www.cei.uni-muenchen.de/. [Accessed 1 November 2015].

DigiPal. (2010–14). *Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic*. London: King's College. Available at http://www.digipal.eu. [Accessed 1 November 2015].

Exon. (2015–). *The Conqueror's Commissioners: Unlocking the Domesday Survey of SW England*. London: King's College. Available at http://www.exondomesday.ac.uk. [Accessed 1 November 2015].

PUhMA. *Papsturkunden des hohen Mittelalters*. Erlangen: University of Erlangen-Nuremberg. https://www5.cs.fau.de/de/papsturkunden-des-hohen-mittelalters/startseite/. [Accessed 3 March 2016].

MoA (2015–). *Models of Authority: Scottish Charters, c.1100–c.1250*. London: King's College. Available at http://www.modelsofauthority.ac.uk. [Accessed 3 March 2016].

MoM-CA. *Source Code of the Monasterium Collaborative Archive*. https://github.com/icaruseu/mom-ca. [Accessed 3 March 2016].

Pierazzo, E. and Stokes, P.A. (2011). Putting the text back into context: A codicological approach to manuscript transcription. In Fischer, F., Fritze, C. and Vogeler, G. (eds), *Kodikologie und Paläographie im Digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*. Norderstedt: Books on Demand, pp. 397–430. Available at http://kups.ub.uni-koeln.de/4360/. [Accessed 1 November 2015].

Stokes, P. A., and Noël, G. (2010). Modelling and system design. *Anglo-Saxon Cluster Project Report*. London: King's College. Available at http://www.ascluster.org/techinfo/analysis/system_design.html. [Accessed 1 November 2015].

Stokes, P.A. (2012). Modeling medieval handwriting: A new approach to digital palaeography. In Meister, J.C. *et al.* (eds), *DH2012 Book of Abstracts*. Hamburg: University of Hamburg, pp. 382–5. Available at http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-medieval-handwriting-a-new-approach-to-digital-palaeography. [Accessed 1 November 2015].

Stokes, P.A., Brookes, S., Noël, G., Buomprisco, G., Marques de Matos, D., and Watson, M. (2014) The DigiPal framework for script and image. In *Digital Humanities 2014 Book of Abstracts*. Lausanne: 512–14. Available at http://dharchive.org/paper/DH2014/Poster-193.xml. Poster available at http://www.digipal.eu/media/uploads/uploads/images/blog_posts/2014/DH2014%20Poster.pdf. [Accessed 1 November 2015].

Stokes, P. A. (2015). The problem of digital dating: A model for uncertainty in medieval documents. In *Digital Humanities 2015 Book of Abstracts*. Sydney: University of Western Sydney. Available at http://dh2015.org/abstracts/. [Accessed 1 November 2015].

Stutzmann, D. (2012). Modélisation des signes graphiques (1). *Écriture médiévale et numérique*. Available at http://oriflamms.hypotheses.org/921. [Accessed 1 November 2015].

Stutzmann, D. (2013). Ontologie des formes et encodage des textes manuscrits médiévaux: Le projet ORIFLAMMS. *Document númerique* 16(3): 81–95.

TEI: Text Encoding Initiative (2015). Representation of primary sources. *P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.9.1 revision 46ac023. Available at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html. [Accessed 1 November 2015].

Webber, T. (2015). The handwriting of Scottish charters 1100–1250 in the National Library of Scotland. *Models of Authority: Scottish Charters and the Emergence of Government 1100–1250*. King's College London. http://www.modelsofauthority.ac.uk/blog/handwriting/. [Accessed 3 March 2016].

Diachronic changes of the Russian Presidential Addresses to the Federal Assembly: From the perspective of archetype key words

Mao Sugiyama
sugiyama.mao0420@gmail.com
Osaka University, Japan

The purpose of this study is to investigate diachronic changes of the Russian Presidential Addresses to the Federal Assembly (RPAFA) given by Russian Presidents, Boris Nikolayevich Yeltsin, Vladimir Vladimirovich Putin and Dmitrii Anatolievich Medvedev. This study proposes political archetype key words (PAKW) and provides a fine-grained classification of presidential speeches.

Russian presidents have given a number of annual speeches. Among them, the RPAFA, which started in 1994, has drawn text-miners' attention. In one recent study, Jashin (2010), argues the importance of "political archetype key words" (PAKW) in the RPAFA through one text from each president: Yeltsin in 1996, Putin in 2000 and Medvedev in 2008. While thought provoking, Jashin's observation has two fatal issues: (i) his observation is based on raw frequencies of each lexical item, and (ii) the speech texts that he uses are arbitrarily limited. That is, he only uses three of the speeches, one from each president, without providing reasons for not using all the texts. Thus, the tendencies observed by Jashin cannot be straightforwardly generalized to tendencies of the presidents; Rather, they may be tendencies of the year. Though the notion of the PAKWs is useful, a further investigation of the PAKWs is needed.

In order to overcome the insufficiencies in Jashin's study, the data set must be reexamined. The data used in this study includes the following:

1. the full set of RPAFA speeches (22 texts in total)
2. the 30 most frequent content words as PAKWs

Importantly, this study intentionally excludes function words because only content words can convey political messages.

The result of the correspondence analysis using a frequency matrix of the PAKWs in each RPAFA file is given in Figure 1 and Figure 2.

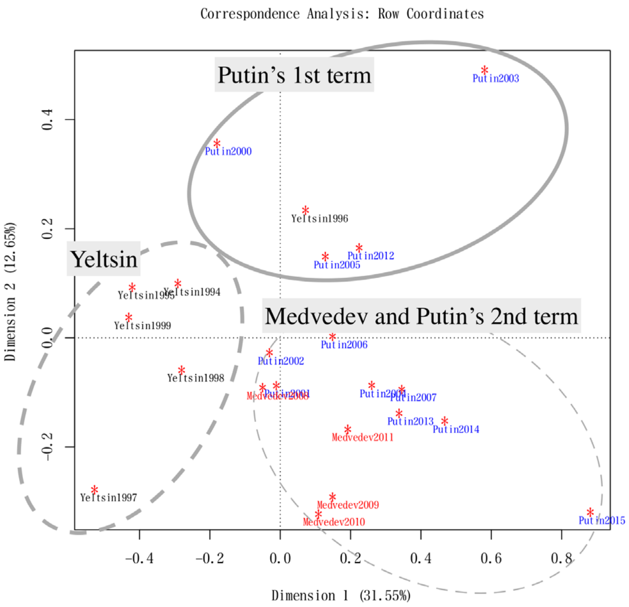


Figure 1 The textual distribution

According to Figure 1, we can roughly classify the speeches into three large categories: Yeltsin on the leftmost circle in a bold dotted line (hereafter, Y dimension); Putin's first term circled with a solid line (henceforth, P dimension); and Medvedev and Putin's second term traced by a fine dotted circle (thence, MP dimension).

Figure 2 shows PAKWs that each text frequently uses. Words observed around the Y dimension are highly relevant to national reconstruction, such as *власти* 'authority-GEN', *государственной* 'state-GEN-FEM', *органов* 'organ-GEN-PL', *деятельности*'activity-GEN'. Putin in the P dimension prefers to use words as like in Yeltsin, but Putin tends to use words related to *Russia* such as *Россия* 'Russian-NOM', *российской* 'Russian-GEN-FEM' in his first term. In the MP dimension, the figure shows that the presidents frequently used words related to the life of citizen like *жизни*'life-GEN', *люди* 'people-GEN-PL', экономика 'economy-GEN', *сфера* 'sphere-PREP' and правительство 'government-subject case'.

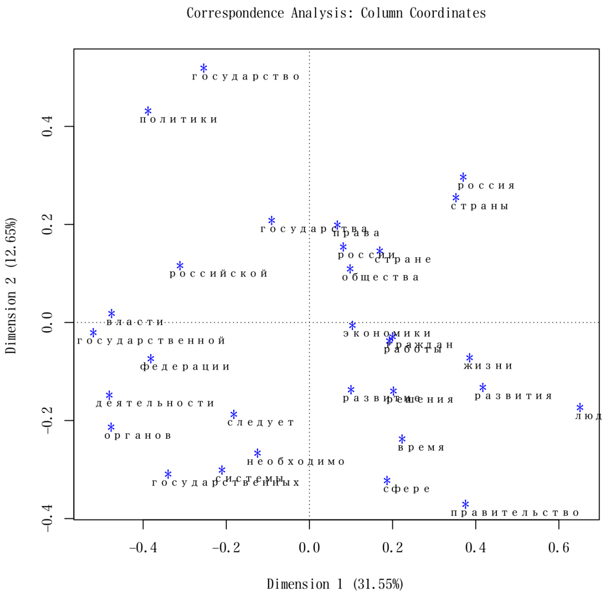


Figure 2 The distribution of PAKWs

The observation based on the distributions of the PAKWs suggests that the classification based on the correspondence analysis in Figure 1 is attributed to the historical backdrop. Let us consider Figure 2, which corresponds to the result of Figure 1. In the Y dimension, Yeltsin frequently uses words related to *state*. In other words, he expresses concern through words related to 'state'. This is primarily because the Russian government was in confusion due to the collapse of the Soviet Union, and Yeltsin's primary policy was to construct a democratic political system. A lexical overlap in the Y and P dimensions indicates that Putin inherited Yeltsin's policy, but the use of *Russia* refers to a change of approach to reconstruct Russia from Yeltsin's administration. The MP dimension includes the PAKWs related to non-authorized citizen or nation. This is attributed to development of the country. That is, after enhancement of the Russian political infrastructure supported by Yeltsin and Putin's first term, the primary political attention has shifted to improvement of social-infrastructure in Medvedev and Putin's second term.

To sum up the investigation from Figure 3 and Figure 4, the use of PAKWs in the RPAFA provides political attitudes of the presidents. That is, the 30 most frequent content words show background political issues of the time. This study concludes a possible implication in supporting the political discourse analysis theories.

Bibliography

Ishino, T. (2014). *Putin daitouryou nenjikyousho enzetu* (2014.12.4) [President Putin annual State of the Union speech: 2014. 12. 14], Russia -related memo 109

Ueno, T. (2009). *Medvedev daitouryou no seijikaikaku – 2008 nen do kyoushoenzetu ni okeru seijikaikaku teian* [Political reform of President Medvedev: On political reform proposals in the 2008 State of the Union speech'] , International Issues 580, p. 4-15.

Кузнецов, В. И. (2014). *Формирование политического курса президента Б. Н. Ельцина в посланиях федеральному реализации* [Formation of president B. N. Yeltsin's policy in the Messages to the Federal Assembly of the Russian Federation (For the period 1994-1999) and the problems of its realization.] Политология No. 3 p. 1-11.

Яшин, В. Н. (2010). *Анализ системы архетипических ключевых слов современной Российской политической речи*[The analysis of the system of archetype key words in Russian modern political speech], ИЗВЕСТИЯ ВОЛГОГРАДСКОГО ГОСУДАРСТВЕННОГО ПЕДАГОГИЧЕСКОГО УНИВЕРСИТЕТА No. 2, p. 120-124.

Imao, Y. (2015). CasualConc (Version 2.0) [Computer Software]. Osaka, Japan: Osaka University. Available from https://sites.google.com/site/casualconc/(accessed 3 March 2016).

Ruthenian Metrica: technological aspects of the electronic publication

Yury Anatoliyovych Svyatets
devotee@i.ua
Oles Honchar Dnipropetrovsk National University, Ukraine

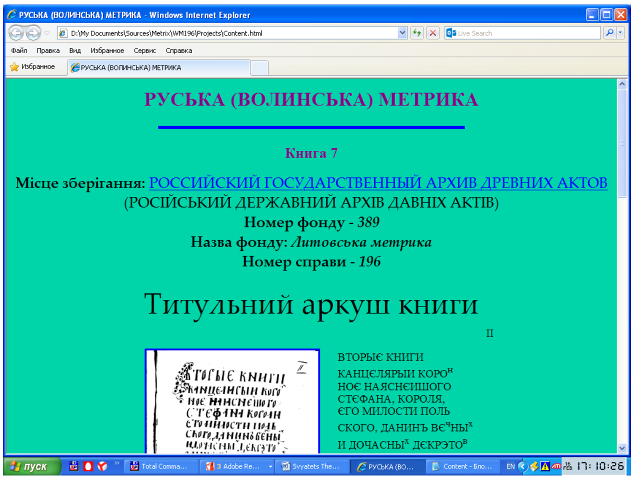
Ruthenian (Volhynian) Metrica is a complex of books of acts of ruthenian (from Ruthenia) series of Crown Chancery of the Polish-Lithuanian Commonwealth. 29 books in which more than 3,5 thousand acts are entered were a part of the Metrica. The main part of documents is written in the Cyrillic letter of Ruthenian language. Other part of acts is written down by the Latin letter on Latin or Polish. Copies of confirmations (privilea) and the decision of royal court are entered in books (decrees). Therefore Ruthenian Metrica consists of books of two types: blotters and books of decrees. Each act has the original name. All acts are rather precisely dated. Documents of the Metrica can be divided into two categories conditionally: simple and compound. Compound documents usually contain literally entered copies of other acts (vidimus) in the structure. Usually entered as private-law acts (for example, wills or donative), and decisions of local courts.

For the purpose of ensuring broad access for historians to documents of the Ruthenian Metrica on department of history of the Oles Honchar Dnipropetrovsk National University (Ukraine) the project on creation of the Web publication of such complex of sources is carried out. Implementation of the project assumes creation of an online-resource in which will various opportunities of presentation and the analysis of documents are integrated.

The concept of the Web version of a complex of acts of this book provides possibility of simultaneous display as the digitized images of pages, and their paginal transliteration in a text format. Access to all acts of the book is provided from the initial page where the general register is placed (fig. 1). The register allows to open its PDF version according to the name of the document or at the choice of number of a leaf to pass to the corresponding Web version (fig. 2). In each case the accompanying information on a place of storage of the original of the act is submitted. It is possible to see documents both in a hand-written form, and in the transcribed (fig. 3). As the cursive writing still kept portable letters, and in the electronic text means of HTML transferred such feature of the letter. It is possible to increase the digitized image of the page of the document in a separate window, having activated the reference (fig. 4). The hypertext structure is comfortable also that regarding the electronic text it is possible to bring later references to the connected documents of other books of Ruthenian, Lithuanian or Crown Metrica, magistrate or territorial government or judicial books both archival, and published. Similar additions can concern a personnel, toponyms or hydronyms, groups of documents, special terminology, etc. For transition between pages of the book the corresponding interactive references in a type of numbers of sheets are provided in the top and lower part of the displayed document. The user can always return to the general register of acts of the book from any looked-through page according to the universal reference "Register". The marking of text part on the basis of the TEI standard allowing to process the free text is in the long term planned.

At this stage quality of images of pages of the book rather low in connection with a digitizing of a monochrome microfilm on which they were finished shooting as a backup copy of storage more than three decades ago. Respectively esthetic registration of insufficient level. However in this case images carry out rather confirmation function, allowing the user, to get acquainted with paleography of acts and to make sure of correctness of the transcribed text and if necessary to offer the specifications. Replacement of images on better is defined not so much technological, how many by legal and financial factors.

This project is interesting also to that it has also didactic value. Such form of submission of documents carries out functions as visual aid in which various samples of paleography of a Ruthenian cursive writing and the corresponding transliteration, and the simulator for creation of the Web version of the case of hand-written sources are presented.



РУСЬКА (ВОЛИНСЬКА) МЕТРИКА

Книга 7

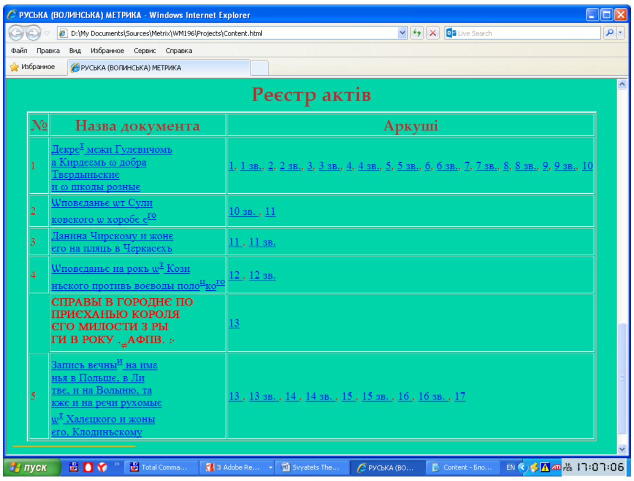
Місце зберігання: РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ АРХИВ ДРЕВНИХ АКТОВ
(РОСІЙСЬКИЙ ДЕРЖАВНИЙ АРХІВ ДАВНІХ АКТІВ)
Номер фонду - ***389***
Назва фонду: ***Литовська метрика***
Номер справи - ***196***

Титульний аркуш книги

II

ВТОРЫЄ КНИГИ
КАНЦЄЛЯРЫИ КОРО$^{\text{Н}}$
НОЄ НАЯСНЄИШОГО
СТЄФАНА, КОРОЛЯ,
ЄГО МИЛОСТИ ПОЛЬ
СКОГО, ДАНИНЪ ВЄ$^{\text{Ч}}$НЫ$^{\text{Х}}$
И ДОЧАСНЫ$^{\text{Х}}$ ДЄКРЭТО$^{\text{В}}$

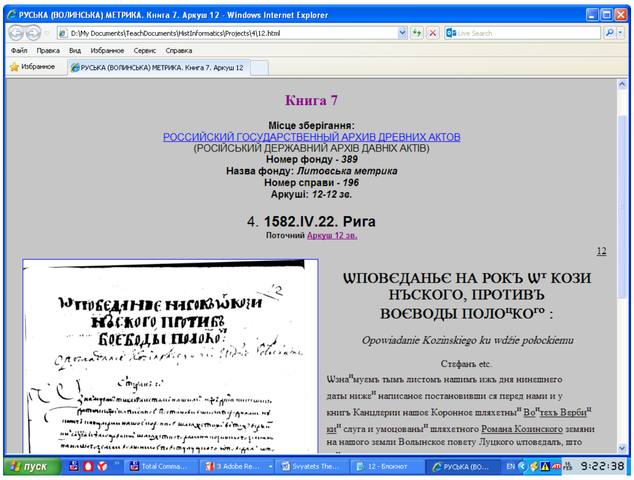
Fig. 1. Initial Site for Access to Acts of the Ruthenian metrica



Реєстр актів

№	Назва документа	Аркуші
1	Декре$^{\text{т}}$ межи Гулевичомъ а Кирдеємъ ѡ добра Твердыньскиє и ѡ шкоды розныє	1, 1 зв., 2, 2 зв., 3, 3 зв., 4, 4 зв., 5, 5 зв., 6, 6 зв., 7, 7 зв., 8, 8 зв., 9, 9 зв., 10
2	Ѡповєданьє ѡт Сули ковского ѡ хоробє є$^{\text{го}}$	10 зв. , 11
3	Данина Чирскому и жонє єго на пляць в Чєркасєхъ	11 , 11 зв.
4	Ѡповєданьє на рокъ ѡ$^{\text{т}}$ Кози нъского противъ воєводы поло$^{\text{ц}}$ко$^{\text{го}}$	12 , 12 зв.
	СПРАВЫ В ГОРОДНЄ ПО ПРИЄХАНЬЮ КОРОЛЯ ЄГО МИЛОСТИ З РЫ ГИ В РОКУ .$_{\neq}$АФПВ. :-	13
5	Запись вєчны$^{\text{и}}$ на имє нья в Польщє, в Ли твє, и на Волыню, та кжє и на рєчи рухомыє ѡ$^{\text{т}}$ Халєцкого и жоны єго, Клодиньскому	13 , 13 зв. , 14 , 14 зв. , 15 , 15 зв. , 16 , 16 зв. , 17

Fig. 2. Ruthenian Metrica Acts Registry


РУСЬКА (ВОЛИНСЬКА) МЕТРИКА. Книга 7. Аркуш 12 - Windows Internet Explorer
D:\My Documents\TeachDocuments\HistInformatics\Projects\4\12.html
Live Search
Файл
Правка
Вид
Избранное
Сервис
Справка
Избранное
РУСЬКА (ВОЛИНСЬКА) МЕТРИКА. Книга 7. Аркуш 12
пуск
Total Comma...
3 Adobe Re...
Svyatets The...
12 - Блокнот
РУСЬКА (ВО...
EN
9:22:38

Книга 7

Місце зберігання:
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ АРХИВ ДРЕВНИХ АКТОВ
(РОСІЙСЬКИЙ ДЕРЖАВНИЙ АРХІВ ДАВНІХ АКТІВ)
Номер фонду - ***389***
Назва фонду: ***Литовська метрика***
Номер справи - ***196***
Аркуші: ***12-12 зв.***

4. **1582.IV.22. Рига**
Поточний Аркуш 12 зв.

12

ѠПОВЄДАНЬЄ НА РОКЪ Ѡ$^{\text{Т}}$ КОЗИ
НЪСКОГО, ПРОТИВЪ
ВОЄВОДЫ ПОЛО$^{\text{Ч}}$КО$^{\text{ГО}}$:

Opowiadanie Kozinskiego ku wdżie połockiemu

Стєфанъ etc.

Ѡзна$^{\text{и}}$муємъ тымъ листомъ нашимъ ижъ дня нинєшнєго даты нижє$^{\text{и}}$ написаноє постановивши ся пєрєд нами и у книгъ Канцлєрии нашоє Коронноє шляхєтны$^{\text{и}}$ Во$^{\text{и}}$тєхъ Вєрби$^{\text{ц}}$ ки$^{\text{и}}$ слуга и умоцованы$^{\text{и}}$ шляхєтного Романа Козинского зємяни на нашого зємли Волынскоє повєту Луцкого ѡповєдаль, што

Fig. 3. An Example of Act Web Page

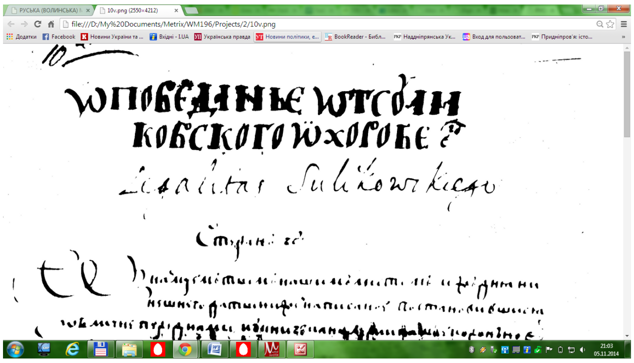


Fig. 4. Increased Act Page

Bibliography

TEI P5: Guidelines for Electronic Text Encoding and Interchange by the TEI Consortium. Originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL Text Encoding Initiative Now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium. The TEI Consortium. Version 2.7.0. Last updated on 16th September 2014,revision 13036. Text Encoding Initiative Consortium, 2014.

The Ruthenian (Volhynian) Metrica. Registers of Polish Crown Chancery Documents Addressed to Ukrainian Lands (Palatinates of Volhynia, Kyiv, Bratslav, and Chernihiv) 1569-1673 with an Introduction by Patricia Kennedy Grimsted. Kyiv, 2002.

Regional Digital Humanities Consortia: An Emerging Formalization of Informal Network Ties?

John Christopher Theibault
jtheibault@gmail.com
PhillyDH

The infrastructure supporting digital humanities work has evolved in a sometimes convoluted pattern of highly resourced large scale initiatives and small scale informal initiatives that lead to durable models that then scale up. Prime examples of informal structures becoming increasingly formalized until they become important features of the digital humanities infrastructure include CenterNet and THATCamp. Understanding such emerging informal initiatives and charting their development to scale is important for determining the essential ingredients of digital humanities infrastructure.

This poster presentation will identify and explore a new informal infrastructure for digital humanities work that has emerged in the last five years: Regional consortia of digital humanities practitioners. These regional consortia bring together people from a range of institutions within a defined geographical area. They can be distinguished from, on the one hand, state and national digital humanities groups that organize conferences and edit journals and require paid membership, as well as digital humanities centers located in a single institution or formally constituted groups with explicit criteria for admission, and, on the other hand, not visibly organized interactions in active digital humanities regions, even if those interactions are frequent in practice. Paris and London might serve as prominent examples of the latter, as neither has developed an informal network despite being highly active digital humanities practitioners. In fact, a central question about the presence of regional consortia is what need is being met by taking on an organized form at all, as opposed to just "getting together"? It has been frequently noted that digital humanities infrastructure is highly unevenly distributed. Will successful examples of regional consortia in some regions inspire or provide a framework for a widening support network in underserved regions? Or will it lead to consolidation of already established digital humanities communities?

I identify eleven regional consortia that have organized themselves with enough visibility to enable study (ten in the United States and one in Europe): DH SoCal, PhillyDH, NYCDH, Boston DH, Virginia DHC, TexasDH, Florida DH, SFBay DH, Detroit DH, Keystone DH, and DigHum Berlin. The framework is heavily tilted towards organizations in the US, and it is possible that there are comparable organizations in other countries that I have failed to uncover. As far as I can determine, each of these groups emerged more or less spontaneously from local initiatives, rather than being directly modeled on one another. I am not including DHBenelux or DH Nordic in my analysis though they share some of the same markers of spontaneous creation "from below," because both have formally become affiliates of EADH and thus have taken on characteristics of national organizations. Many of them (though not all) share a close association with THATCamp. Either they came into being as part of the process of organizing a THATCamp or they were result of a THATCamp session. At a minimum, they exist virtually, as either a website, a twitter feed, or a listserv, or often all three. Usually, they also have face to face meetings of some kind. While most regional consortia have emerged in regions with at least one significant digital humanities center located at a major university, one of their most conspicuous features is that they are all expressly not housed at a single institution and are open to anyone interested in digital work in the region. Indeed, most appear to encourage outreach to under-resourced local cultural heritage institutions as part of their activities. Costs are kept very low, though some regional consortia have explicit institutional support from universities in the region, while others rely strictly on members' support, drawing on institutional resources that individuals control.

The earliest regional consortium to establish an online presence, DH SoCal, first posted in March, 2010. Because such regional consortia are so new, they are still in the process of defining their missions and organizational structures. In almost all instances, an important reason for reaching out across institutional lines in a region is a perception that the range of skills/technical expertise required for digital projects is too great for any one institution to have them all. Regional consortia become a forum for sharing expertise. At the same time, as noted above, regional consortia act as evangelizers for digital projects, to encourage local cultural heritage organizations to embrace digital work. So the other primary function of consortia is to publicize local digital projects to a wider public. Membership in these consortia vary greatly in terms of the distribution between students, faculty, librarians, and representatives of cultural heritage organizations. This poster will compare and contrast the membership, statements of purpose, and activities of these various consortia to see if there are core features that may help future groups organize.

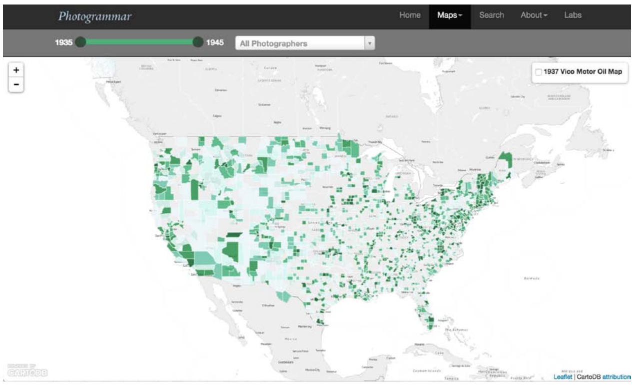
Photogrammar

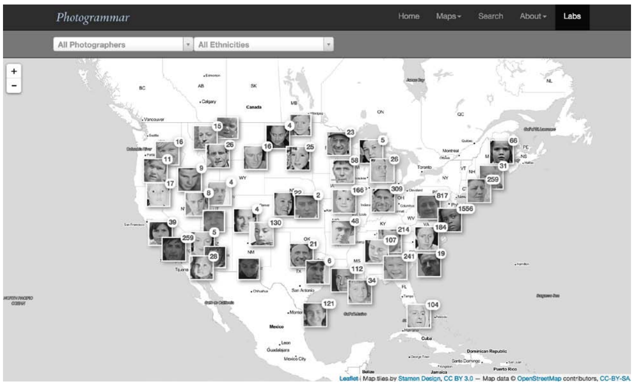
Lauren Tilton
lauren.tilton@yale.edu
Yale University, United States of America

Taylor Arnold
Taylor.Arnold@yale.edu
Yale University, United States of America

Peter Leonard
Peter.Leonard@yale.edu
Yale University, United States of America

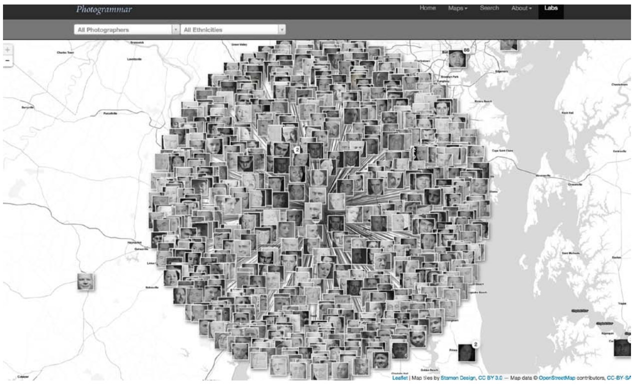
The poster will focus on Photogrammar (photogrammar.yale.edu), a web-based platform for organizing, searching, and visualizing the 170,000 photographs from 1935 to 1945 in the United State's Farm Security Administration and Office of War Information (FSA-OWI) Archive. Our poster session will include two parts – a physical poster and demonstrations of how to use Photogrammar. The goal is to highlight how we used digital humanities methods such as text, image and spatial analysis to increase discoverability of a photograph archive while also changing humanities scholarship related to the FSA-OWI.



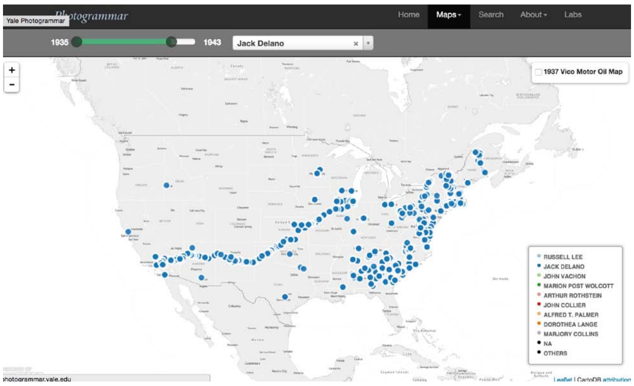

Photogrammar
Home
Maps
Search
About
Labs
All Photographers
All Ethnicities
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors, CC-BY-SA

For the poster, we will first outline how Photogrammar is changing scholarship in media studies, visual culture studies and 20th century United States cultural history. In order to build support for and to justify government programs during the Great Depression and World War II, the FSA-OWI set out to document America and the successful administration of government services. They produced some of the most iconic images of the era and employed prominent documentary photographers such as Arthur Rothstein, Dorothea Lange, Gordon Parks, and Walker Evans, all of whom shaped the visual culture of the era both in its moment and in American memory. Unit photographers were sent across the country. The negatives were sent to Washington, DC. 170,000 negatives were collected and for decades, scholars have argued that FSA-- OWI archive is a collection about rural poverty in the American south and Dust Bowl. Photogrammar shows that mapping the photographs challenges decades of scholarship on one of the most prominent visual culture archives in the United States.

The photographers took pictures across the nation, which leads to new questions about the breadth and depth of the archive and goals of the federal government during the era. Questions include why the United States federal government sought these images and how individual photographers' ways of seeing impacted the archive. We will then turn to the methods used to reframe and visualize the photographs that are changing scholarship on and the discoverability of the FSA--OWI collection. We will focus on three techniques: automated geo-referencing and CartoDB to map the photographs, TF--IDF and cosine similarity to surface related photographs based on their captions, and facial recognition software to make the collection searchable by faces (OpenBR).


Photogrammar
Home
Maps
Search
About
Labs
All Photographers
All Ethnicities
+
−
86
2
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors, CC-BY-SA

Along with the poster, we will offer demonstrations of the site. We will highlight the multitude of features on the site such as tracking photographers as they move across the country and how users can explore the collection through the archival system developed by Paul Vanderbilt in the 1940s to organize the collection. We will also share new Labs that are in development in order to receive feedback from the international DH community.



We are also excited to share with participants of DH2016 how we built Photogrammar and to speak with them about how they could apply these methods to their archives and humanities data. In addition, we are eager to learn from participants about additional techniques that might augment Photogrammar.

Automatic quotation detection in Russian nonfiction texts

Nataliya Tyshkevich
natalie.tysh@gmail.com
National Research University Higher School of Economics, Moscow, Russia, Russian Federation

1 Introduction

We present work in progress to automatically extract quotation constructions. We claim that it is possible to infer markers that introduce quotation even for the languages without grammaticalized reported speech forms. The object of our study is not the quotation itself, but the canonical context of quotation. We rely on the concept of context as a term with a "hole" (Gabbay and Lengrand, 2008). The context with a quotation as filling for the "hole" has formal features, which we consider as parameters of quotation canon in terms of the canonical typology (Corbett et. al, 2012). We specialize in direct written quotations, which can be distinguished by the reader exactly because of its heterogenity.

We used semi-supervised machine learning techniques to discover contribution of individual predictors to the quotation distribution in Russian nonfiction texts and to build a working prototype of quotation classifier. The classifier does not need any particular text corpora on the final stage, using it only as training data. It predicts probability of sentence to contain a quotation with accuracy of 85%. The objective of the research is to create a machine-learned model that would distinguish between citations and non-citations and arrange the quotation boundaries. The paper presents such a model trained on Russian data. As a result importance of different contextual features can be measured.

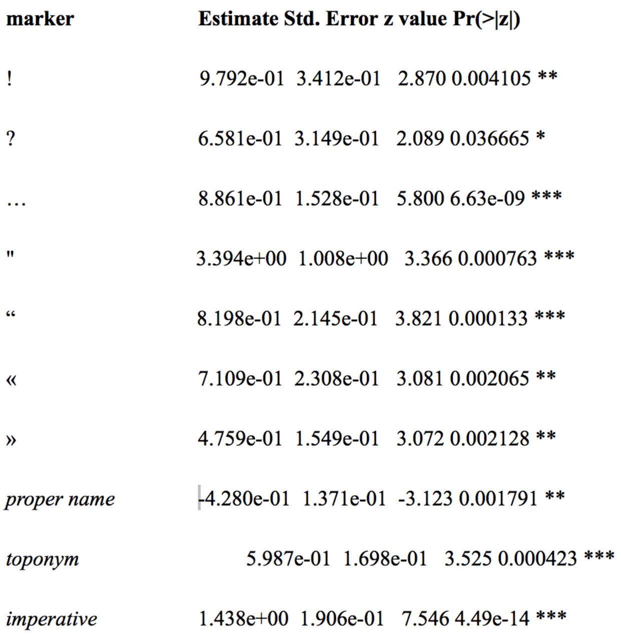
Our model can be used both by scholars and by ordinary readers, unaware of intertextuality issues, and our achievements are undoubtedly useful for a wider class of problems, identifying heterogeneous fragments in the text.

2 Dataset

The material for the training sample is based on the dictionary of Russian literary quotations (Dushenko, 2005) and on the corpora of Russian nonfiction texts from the web-project "Zhurnal'ny Zal" (ZZ) (http://magazines.russ.ru/). We formed 24761 search queries and extracted approximately 1200 sentences, containing quotations from the dictionary and 1200 sentences without any of them. Semi-automatic filtering of false quotative sentences and correcting of quotations boundaries gave us the training corpora of 1000 quotation sentences with the correct boundaries, 1000 sentences without any quotations and 1000 examples of quotative sentences with false borders. At the machine learning stage we divided it to two complementary sets for training and testing respectively.

3 Features

We define overt quotation markers as the most relevant context features for identifying the quote. We examined these features in the tagged corpora and built a logistic regression model, providing us with the most significant markers and their combinations (Fig. 1):



marker	Estimate	Std. Error	z value	Pr(>|z|)	
!	9.792e-01	3.412e-01	2.870	0.004105	**
?	6.581e-01	3.149e-01	2.089	0.036665	*
…	8.861e-01	1.528e-01	5.800	6.63e-09	***
"	3.394e+00	1.008e+00	3.366	0.000763	***
“	8.198e-01	2.145e-01	3.821	0.000133	***
«	7.109e-01	2.308e-01	3.081	0.002065	**
»	4.759e-01	1.549e-01	3.072	0.002128	**
proper name	-4.280e-01	1.371e-01	-3.123	0.001791	**
toponym	5.987e-01	1.698e-01	3.525	0.000423	***
imperative	1.438e+00	1.906e-01	7.546	4.49e-14	***

Figure 1.

That allowed us to exclude the most unimportant features from our final model.

4 Model

The main learning property of the Binary Machine was distinguishing between the presence and absence of a particular marker in certain parts of a sentence. Thus, we built two classifiers:

1) dividing the examples into quotative and non-quotative;

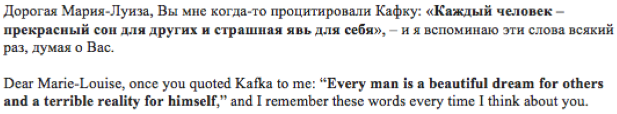
2) pointing out the quotation boundaries.

The final response was based on the cumulative response of the two classifiers. The first machine would give a positive response with a probability 0.7 and more, whereas the second classifier would activate choosing the most relevant hypothesis out of the potential set of quotation boundaries. Support Vector Machines method is commonly applied for classification tasks. We used a modification of this method, Support Vector Classification, because it gives a best fit to our binary data type.

5 Results

The two classifiers have shown positive results with the quotation classifier accuracy of 0.86 and borders classifier accuracy of 0.83.

The greatest confidence in the quotativeness of the fragment results from the presence of the quotation marks. Without them the precision of the evaluations decreases, but not critically. What is important is that our model is not oriented to the quotation marks, the most obvious marker, it analyses all the relevant features in the sentence. For example, for the quotative sentence with overt quotation marks (Fig. 2) probability was estimated as 99%, and the same sentence with cut quotation marks was evaluated as 98% quotative.



Дорогая Мария-Луиза, Вы мне когда-то процитировали Кафку: «**Каждый человек – прекрасный сон для других и страшная явь для себя**», – и я вспоминаю эти слова всякий раз, думая о Вас.

Dear Marie-Louise, once you quoted Kafka to me: “**Every man is a beautiful dream for others and a terrible reality for himself**,” and I remember these words every time I think about you.

Figure 2.

The final version of the program allows one to automatically mark quotes in an untagged text with the permissible share of errors. It works best with nonfiction texts and operates either plain text, or group of texts, or one sentence. Also it can be re-learned on the following set of interchangeable input data for training sample: the set of relevant quotation parameters, a list of standard quotations and group of texts for context-mining. We expect this method to be rather flexible and applicable to other text corpora.

Bibliography

Corbett et al. (2012). Canonical morphosyntactic features. In Dunstan Brown, Marina Chumakina and Greville G. Corbett (eds.), Canonical morphology and syntax , 48-65. Oxford: Oxford University Press.

Gabbay (2008). Gabbay M. J., Lengrand S. The λ-context calculus //Electronic Notes in Theoretical Computer Science. – T. 196. – C. 19-35.

Dushenko (2015). Citaty iz russkoj literatury: 5200 citat ot «Slova o polku.» do nashih dnej.

ZZ. “Zhurnal'nyj zal”, http://magazines.russ.ru/. (accessed on 2016–03–06).

Project COLEM for CREATE (University of Amsterdam) Adapting NPL-Tools for Creating an Orthographic Layer for Early Modern Dutch Texts

Wouter van Elburg
w.m.vanelburg@uva.nl
University of Amsterdam, Netherlands, The

Tessa Wijckmans
t.m.wijckmans@uva.nl
University of Amsterdam, Netherlands, The

The seventeenth century Dutch language did not know a standardized spelling. Because of this, many different spelling variants of the same word (i.e. *ik* and *ick*, both meaning *I*) coexist. This may largely be the result of an author's own preferences, background or the book printer he used. Software to standardize, analyze or process texts is mostly developed for modern Dutch, so processing historical texts with natural language processing (NPL) tools or analysing such texts on stylometric aspects is problematic. Due to the orthographic variation of historical texts, software will not always recognize words that have the same meaning, but different spelling.

For project COLEM we want to normalize spelling differences by having digital re-speller tools form a standardized spelling variant, that could help software better understand the texts. We investigate the possibilities to provide the original text with an orthographic layer containing the normalized words. In this way, the original texts and its morfo-syntactic information are still accessible and it will be possible to search both the original text and the layer. This, for instance, will ease research to language evolution.

The Java software VARD2 (Baron 2011; Baron and Rayson, 2008) seems to be a useful tool for this purpose. This tool was originally created to normalize old English texts, but can be adjusted to other languages. VARD2 will compare the words in the input text with an incorporated, but easily adaptable wordlist (a .txt-file). We replaced the default wordlist with a Dutch lexicon and we trained VARD2 on texts of two different seventeenth century Dutch authors: Simon de Vries and Gotfried van Broekhuizen. Texts of these authors are characterized by a significant different orthography and this could therefore help us to train normalizers and test them on different spelling forms existing within the Early Modern Dutch language.

We trained VARD2 by replacing the variants in the historical texts with a normalization. VARD2 will present suggestions for normalization by using four methods of which just one is language dependent (a modified version of the Soundex phonetic matching algorithm that is based on English phonemes). Two other methods, that of letter rules and that of known variants, we adapted by replacing the default .txt-files 'letter rules' and 'known variants'. The last method, a normalized Levenshtein Distance, does not need to be adapted, since it is language independent.

In this presentation we will show the performances of our trained VARD2 tool. We will focus on a specific amount of problems the tools run into, like uncommon words, clitics and combined words. We will also investigate the possibilities of the Norma tool (Bollman, 2012) and TICCLops (Reynaert, 2014). By comparing the results that various tools offer, we will decide what tool(s) is/are most successful in dealing with these problems. This could give us ideas for follow-up research or the development of tools for normalizing Early Modern Dutch, but probably also for normalizing other languages with unstable orthographies.

Bibliography

Baron, A. (2011). *Dealing with spelling variation in Early Modern English texts.* Ph.D. thesis, University of Lancaster.

Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. *Proceedings of Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University, May 2008.

Bollmann, M. (2012). (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In Mambrini, F., Passarotti M. and Sporleder C. (eds), *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ARCH-2)*. Lisbon, Portugal, pp. 3-14.

Reynaert, M.W.C. (2014). Synergy of Nederlab and PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland, 2014.

Notes

[1] It is not known who exactly was responsible for the spelling as it was printed. But due to there being examples of texts printed by the same printer that use radically different spelling forms it is plausible that the author of the text was responsible for the spelling.

[2] VARD is an acronym for Variant Detector.

[3] Norma is written in C++11, though bindings for Python are provided as well.

[4] TICCLops v.0.2(Text Induced Corpus Clean-up online processing system) is a web application (offered in a JavaScript interface) that is intended to detect and correct typographical errors and OCR (optical character recognition) errors in text. It is usable for every language, since it bases its replacements on the input corpus by making Most Frequent Words-list. However, TICCLops is probably less usable for providing a text with an annotation layer, because the replacements take place in the texts itself, without preserving an original version of the text.

Reconstructing Past Teaching

Demmy Verbeke
demmy.verbeke@kuleuven.be
KU Leuven, Belgium

Sam Alloing
sam.alloing@kuleuven.be
KU Leuven, Belgium

Luc Lannoy
luc.lannoy@kuleuven.be
KU Leuven, Belgium

Matthias Meirlaen
matthias.meirlaen@kuleuven.be
KU Leuven, Belgium

Bruno Vandermeulen
bruno.vandermeulen@kuleuven.be
KU Leuven, Belgium

Ilse Neirinck
ilse.neirinck@kuleuven.be
KU Leuven, Belgium

One of the best sources to get an insight in how various subjects (e.g. history, Latin, home economics) were instructed to children and teenage pupils in the past, are the textbooks that were used in various pedagogical systems. They can teach us how knowledge in these various subjects was ordered and presented and how one tried to transfer this knowledge – a fact that is recognized, e.g., in the multifarious work of researchers working on the collections of the National Museum of Education (Rotterdam) or the Georg Eckert Institute for International Textbook Research (Braunschweig). This type of publications, however, is often overlooked in the creation of library collections. The main reason for this is that textbooks, and other connected teaching materials (such as curricula, didactic plates and pedagogical journals), are generally considered to be only relevant for a brief period of time, namely for as long as they are used in a didactic context. They are continually replaced by more recent materials, and only become the object of study again when one wants to reconstruct past teaching.

This poster depicts the efforts at the University Library of the KU Leuven (Belgium) to preserve and prepare such didactic material, created in a Belgian context and thus mostly written in French or Dutch, for present and future research. It also presents how specific corpora of textbooks are catalogued and digitized to facilitate their study, and details the concrete plans to further unlock these didactic sources through more advanced digitization enabling distant reading (starting from an exploratory pilot study focusing on history textbooks). By presenting this case, it illustrates the current results and the future possibilities of the collaboration between researchers and library staff in the context of specific Digital Humanities projects, and documents the ambition of various divisions of the University Library to continue to foster such collaborations (Verbeke, 2014 and Truyen-Verbeke, 2015).

The poster first describes the origins and content of the collection of textbooks and related didactic materials within the Library of Psychology and Educational Sciences, a division of the University Library at KU Leuven. It then details how this unique collection is being catalogued by the staff of the Metadata Unit. This is done according to a specifically developed data model (using the resource management environment in the library software *Alma*) which both ensures maximum compatibility with the general library discovery system (currently *Primo/Limo* at the KU Leuven), as well as the provision of detailed metadata which is needed in the context of an academic study of these teaching materials. The next part discusses the various ways in which this collection is or can be used by researchers, as well as the various selections which have been made in the context of specific research projects (e.g. by focusing on representations of Belgian-Congolese (post) colonial history in Belgian secondary school education since the Second World War, cf. Van Nieuwenhuyse, 2014 and 2015). It details how research projects of this kind determine a corpus of teaching materials, smaller than the entire collection, which are studied in more detail through a close reading of the physical materials, but which also form a priority category of items to be digitized by the Digitization Unit of the University Library. At present, this digitization offers high-end photographic images of the relevant books (using a Qidenus Smart Book Scan, 2 Nikon D800 cameras with 36 million pixels used in a color-calibrated workflow, producing RAW, TIF and JPEG files), but the ambition is to extend this to the creation of a machine-readable textual corpus which enables distant reading as well (with the intention that this distant reading will supplement, not supplant, close reading; see Moretti, 2013 and the varying interpretations of Moretti's intentions in Serlen, 2010, Khadem, 2012, and Ross, 2014). The poster therefore closes with a brief presentation of the OCR (Aletheia, ABBYY Finereader Engine, ocrevalUAtion) and NER (INL Attestation tool, Stanford NER tool with Europeana Newspaper extension) technology used at the University Library of KU Leuven, building on previous experience gathered in the context of a Succeed-project (Alloing-Verbeke, 2014 and Verbeke, 2015).

Bibliography

Alloing, S. and Verbeke, D. (2014). Tools Evaluation: University Library of KU Leuven. http://www.digitisation.eu/blog/tools-evaluation-university-library-ku-leuven (accessed 20 February 2016).

Khadem, A. (2012). Annexing the unread: a close reading of "distant reading", *Neohelicon*, **39**: 409-421.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Ross, S. (2014). In Praise of Overstating the Case: A review of Franco Moretti, Distant Reading (London: Verso, 2013), *Digital Humanities Quarterly*, **8**(1), http://www.digitalhumanities.org/dhq/vol/8/1/000171/000171.html (accessed 20 February 2016).

Serlen, R. (2010). The Distant Future? Reading Franco Moretti, *Literature Compass*, **7**(3): 217-225.

Truyen, F. and Verbeke, D. (2015). The library as a valued partner in Digital Humanities projects: The example of EuropeanaPhotography, *Art Libraries Journal*, **40**(3): 28-33.

Van Nieuwenhuyse, K. (2014). From triumphalism to amnesia. Belgian-Congolese (post)colonial history in Belgian secondary history education curricula and textbooks (1945-1989), *International Journal of Research on History Didactics, History Education, and History Culture*, **35**: 79-100.

Van Nieuwenhuyse, K. (2015). Increasing criticism and perspectivism: Belgian-Congolese (post)colonial history in Belgian secondary history education curricula and textbooks (1990-present), *International Journal of Research on History Didactics, History Education, and History Culture*, **36**: 183-204.

Verbeke, D. (2014). The opportunistic librarian: A Leuven confession, *Digital Humanities 2014: Conference Abstracts*. Lausanne: EPFL and UNIL, pp. 395-396.

Verbeke, D. (2015). Renaissance Studies, Digital Humanities and the Library. *JNR Polaris*, http://www.northernrenaissance.org/renaissance-studies-digital-humanities-and-the-library (accessed 20 February 2016).

Beyond Digital Humanities? Furthering the Exploration of Language Diversity and Pan-European Culture by Means of Transdisciplinary Research Infrastructures: Introducing the new DARIAH CC Science Gateway

Eveline Wandl-Vogt
eveline.wandl-vogt@oeaw.ac.at
Austrian Academy of Sciences, Austrian Centre for Digital Humanities; AT

Roberto Barbera
roberto.barbera@ct.infn.it
Institutio Nationale de Fisica Nucleare; IT

Giuseppe La Rocca
giuseppe.larocca@ct.infn.it
Institutio Nationale de Fisica Nucleare; IT

Antonio Calanducci
antonio.calanducci@ct.infn.it
Institutio Nationale de Fisica Nucleare; IT

Tibor Kalman
tibor.kalman@gwdg.de
Gesellschaft für wissenshaftliche Datenverarbeitung; DE

Thordis Ulfarsdottir
disa@hi.is
The Arni Magnusson Institute for Icelandic Studies; IS

Jozica Skofic
Guzej@zrc-sazu.si
Ran Ramovs Institute of the Slovenian Language ZRC-SAZU; SI

Jadwiga Waniakova
jadwiga.waniak@uj.edu.pl
Institute of Polish Language, Polish Academy of Sciences; PL

Introduction: The Core Project EGI-Engage

This paper introduces into a new Science Gateway, developed in the framework of the European Horizon 2020 project EGI-Engage (Engaging the Research Community towards an Open Science Commons). EGI-Engage

> aims to accelerate the implementation of the Open Science Commons by expanding the capabilities of a European backbone of federated services for compute, storage, data, communication, knowledge and expertise, complementing community-specific capabilities (EGI-Engage).

Within EGI-Engage there are several clusters of research infrastructures, representing different research communities.

The Arts and Humanities are represented by DARIAH, building the DARIAH Competence Centre (DARIAH CC).

In this paper we focus on transdisciplinary collaboration in the framework of eLexicography in cultural context.

In the last decade it becomes almost impossible to image excellent science without support of e-Infrastructures in the sciences. However, although various research areas, such as medicine, chemistry, and physics, mostly depend on the availability of advanced research infrastructure, the area of Arts and Humanities is still not utilizing the available infrastructure at their full potential, especially but not exhaustively on the example of EGI infrastructures.

To overcome this gap, DARIAH CC aims to provide a wider and more efficient access to, and use of, research e-Infrastructures at EGI level, including transnational access as well as joint research and networking for users coming from the areas of the Arts and Humanities.

The first step in achieving this goal is by providing end-user support, at both technical and service level. One way of fulfilling this mission is by providing a workflow-based Science Gateway.

Customising the Long Tail of Science (LToS): Science Gateway and the Semantic Search Engine (SSE) for the Humanities

For the specific research domain of eLexicography, a Science Gateway is adapted and tailored to meet the needs of the user community coming from the field of Arts and Humanities or, in a more general case, from the Social Science and Humanities. Technically the Science Gateway is based on the generic purpose grid and cloud user support environment (gUSE/WS-PGRADE; LPDS [2015]) as well as gLibrary technologies (INFN [2015]).

The new gateway provides access and compute services for data residing in distributed grid and cloud storages, generic applications, as well as computational resources. The usefulness and evaluation of the specific Science Gateway is demonstrated with two specific pilot applications: "Storing and Accessing DARIAH contents on EGI" (SADE), and "Multi-Source Distributed Real-Time Search and Information Retrieval" (SIR). After the end of the EGI DARIAH CC, it results will be contributed to DARIAH-EU through the DARIAH Virtual Competence Centre 1 "eInfrastructures", which will take appropriate means to support and maintain the results.

From the Semantic Search Engine (SSE) you can search in parallel and in more than 100 languages across various Linked Data repositories:

1) the e-Infrastructure Knowledge Base (KB) containing more than 30 million resources belonging to thousands of semantically enriched Open Access Document Repositories and Data Repositories search results are ranked according to the Ranking Web of Repositories.

2) several resources that are essential for the Arts and Humanities SADE use case in context of cultural lexicography. In this poster we focus on Biodiversity and linguistic diversity and show results for European Repositoris dealing especially with this topic, e.g. Europeana, Cultura Italia, Isidore, OpenAgris and Bio2RDF.

Others can simply added using the configuration options. According to the Humanities requirements, further possibilities are at the moment investigated to be included, e.g. AGROVOC, GEONAMES, GERMANET, ISLEX as well as DBPedia. Furthermore, the team is investigating into supporting ALE workflow opportunities.

From the SSE one can right now search for your plant names and get a full list of results according to the EGI-ontology. Figure 1 gives an overview on the science gateway archidecture:

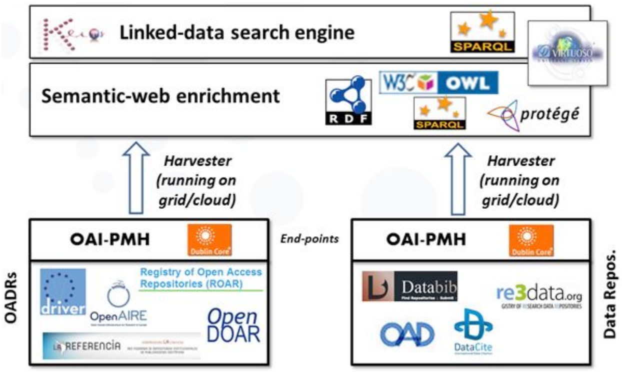
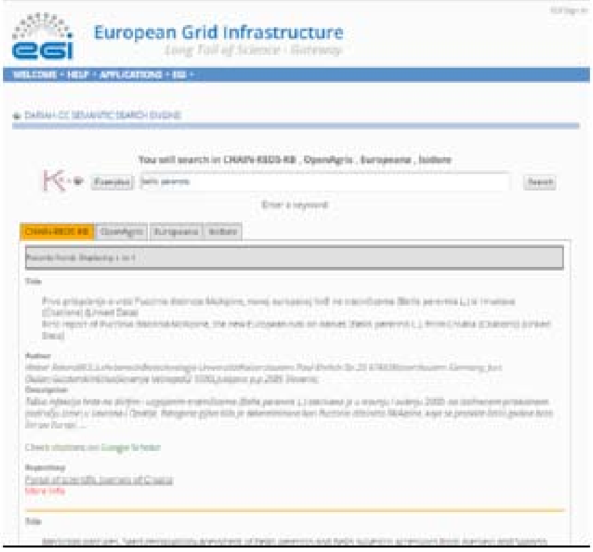


Figure 2 below shows first search results for the search term "bellis perennis":


European Grid Infrastructure
Long Tail of Science · Gateway
WELCOME • HELP • APPLICATIONS • EGI •
DARIAH-CC SEMANTIC SEARCH ENGINE
You will search in CHAIN-REDS-KB , OpenAgris , Europeana , Isidore
Search
Enter a keyword
CHAIN-REDS KB
OpenAgris
Europeana
Isidore
Check citations on Google Scholar
Portal of scientific journals of Croatia
More Info

For each item you can click on:

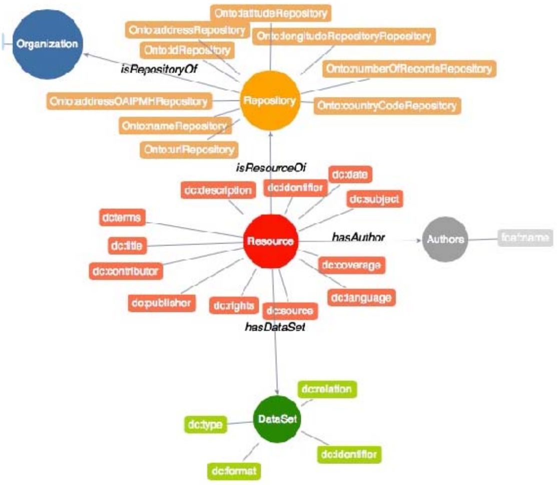
- "More Info" to get more additional information about the dataset

- "Check citations on Google Scholar"

- "Linked Data" to display graphically the semantic connections.
- Access directly the repository where data are stored.

Instructions to access the LTOS SG can be found here.

Figure 3 below introduces the SSE-schema:



Exploring cultural diversity and multilinguality: Supporting ALE and COST ENeL

The certain use case is developed to support Pan-European Lexicography and Lexicology and explore new methods of scholarship. The main use case partners are members of the two European initiatives in the fields, namely for Lexicography the COST action IS 1305 European network for electronic Lexicography (ENeL) and for Lexicology the Atlas Linguarum Europae (ALE).

Both of these initiatives aim to explore cultural diversity against the background of multilingual Europe. Europe, as "common space of knowledge" with a common background for lexicography and with shared lexicographical practices and methodological principles, was to this regards influenced by a common canon of dictionaries and other reference works rather than as independent works (Haß [2010]; Kirkness [2012]) .

Moreover, the purpose of these dictionaries was to contribute to nation-building, which explains their – usually strong – monolingual view. Yet, this does not adequately reflect the common Pan-European linguistic heritage (Munske / Kirkness [1996]; Habermann [1999]) of the languages of Europe, which have been in fact in permanent contact with each other (COST Memorandum of Understanding) .Possibilities to study the Pan-European commonal i ties are explored in the framework of COST ENeL and are to be supported by the science gateway and DARIAH Competenc eCentre.

The Use Case definitely meets user needs: The Memorandum of Understanding for the COST action ENeL [COST MoU, p: 6] refers directly to the lack of common approaches and lacking standards as well as the ambition to "fully explore the possibilities of the digital medium".

The developed Science Gateway is one of the solutions for some of the items aimed to be results of the COST ENeL action Working Group 4 "Lexicography and Lexicology from a Pan-European perspective" (COST / Wandl-Vogt / Nowak) , namely mainly 1) Developing ways in which already existing information from single language dictionaries can be displayed and interlinked to present more adequately their common European heritage – 2) Finding new applications for the very large amount of interconnected dictionary information from the European dictionary portal in the field of digital humanities [COST MoU: 14]. Main topics to be addressed in this paper are forms of scholarship and production of knowledge (Nowotny / Scott / Gibbons [2003]) – a) collaborative – b) explorative – c) interdisciplinary transdisiplinary / in the framework of Open S ci ence.

a) Collaborative: The proposers of the paper discuss how open online collaboration is furthering research on the examples of ALE and COST EneL.

b) Explorative: The authors of the paper introduce how their developed and adapted infrastructure supports explorative scholarship.

c) Interdisiplinary, transdisciplinary, towards Open Science: Finally, they reflect their transdisciplinary collaboration – with background in sciences as well as Humantities – and the added value of it on the one hand; on the other hand they reflect on transforming their workflows towards Open Science Commons.

Conclusion and Outlook: Beyond Digital Humanities?

This paper presents a collaboration beyond Digital Humanities: The contributors introduce into a new science gateway for the Digital Humanities, discuss the main technologies adopted and developed and point at the added value of a collaboration beyond Digital Humanities.

Bibliography

A gricultural Management Standards (2016) AGROVOC: http://aims.fao.org/access-agrovoc (Accessed: 12 March 2016).

Atlas Linguarum Europae (ALE) (2015). http://www.lingv.ro/ALE.html (Accessed: 12 March 201 6).

Bio2RDF (2016). http://pubmed.bio2rdf.org/ (Accessed: 12 March 2016).

Cultura Italiana (2016). http://dati.culturaitalia.it/?locale=it (Accessed: 31 October 2015).

Cybermetrics Lab (2016). Ranking Web of Repositories. http://repositories.webometrics.info/ (Accessed: 31 October 2015).

DBPEdia (2016). http://wiki.dbpedia.org/(Accessed: 12 March 2016).

Europeana (2016). http://www.europeana.eu/portal/ (Accessed: 31 October 2015).

European Cooperation in the Field of Scientific and Technical Research (COST; 2013/05). Memorandum of Understanding for the Implementation of a European Concerted Research Action designated as COST Action IS 1305: European Network of e-Lexicography (EneL). http://w3.cost.eu/fileadmin/domain_files/ISCH/Action_IS1305/mou/IS1305-e.pdf (Accessed: 12 March 2016).

European Cooperation in the Field of Scientific and Technical Research (COST; 2015). COST Action IS 1305: European Network of e-Lexicography (EneL). http://www.cost.eu/COST_Actions/isch/IS1305 (Accessed: 12 March 2016).

European Grid Infrastructure (EGI): https://access.egi.eu (Accessed: 12 March 2016).

European Grid Infrastructure (2015). EGI-Engage. https://www.egi.eu/about/egi-engage (A ccessed: 12 March 2016).

European Grid Infrastructure (2015). L ong Tail of Science (LToS) – Gateway. DARIAH Competence Centre Semantic Search Engine: https://csgf.egi.eu/dariah-sse (Accessed: 31 October 2015).

European Network of eLexicography (2015). www.elexicography.eu (Accessed: 12 March 2016).

European Network of E-Lexicography (ENeL) / Wandl-Vogt, E. , Nowak, K. (2014-): WG4: Lexicography and Lexicology from a Pan-European Perspective. http://www.elexicography.eu/working-groups/working-group-4/wg4-objectives/ (Accessed: 12 March 2016).

Food and Agriculture Organisation of the United Nations (2016): Agris. http://agris.fao.org/openagris/index.do (Accessed: 31 October 2015).

GEONAMES (2016). http://www.geonames.org/ (Accessed: 12 March 2016).

GERMANET (2016). http://www.sfs.uni-tuebingen.de/GermaNet/ (Accessed: 12 March 2016).

Habermann, M. (1999). Latein – Muttersprache Europas, Der Deutschunterricht 3: 2 5-37.

Haß, U. (201 0). Chancen und Perspektiven der historischen Lexikografie des Deutschen. In Lexicographica 2011: 45-62. https://www.unidue.de/imperia/md/content/germanistik/hass/lexicographica_2011_hass_hist_lex_d_deutschen.pdf (Accessed: 12 March 2016).

Instiuto Nationale Fisica Nucleare (INFN) (2015). gLibrary. Digital Libraries on the Grid. https://glibrary.ct.infn.it/glibrary_new/index.php(Accessed: 31 October 2015).

Isidore (2016). http://www.rechercheisidore.fr/(Accessed: 31 October 2015).

Kirkness, A. (2012). Deutsches Wörterbuch von Jakob und Wilhelm Grimm. In Haß, U. (ed .), Große Wörterbücher und Lexika Europas. Europäische Enzyklopädien und Wörterbücher in historischen Porträts . Berlin/Boston, DeGruyter, pp. 211-232.

LPDS, Institute for Computer Sciences and Control, Hungarian Academy of Sciences (2015). Grid and cloud user support environment. http://guse.hu/documentation(Accessed: 31 October 2015).

Munske, H.H. ; Kirkness, A. (1996). Eurolatein. Das griechische und lateinische Erbe in den europäischen Sprachen, Reihe Germanistische Linguistik (RGL), Tübingen, Niemeyer .

Nowotny, H. ; Scott, P.; Gibbons, M . 2003. `Mode 2´ Revisited. The new Production of Knowledge. Minerva 41: 179-194. http://www.uni-klu.ac.at/wiho/downloads/nowotny.pdf . (Accessed: 12 March 2016).

SCI-GaLA (2015). Map of Open Document Repositories (OADRS). http://www.sci-gaia.eu/einfrastructures/ knowledge-base/oadr-map/ (Accessed: Oct, 31st 2015).

The Arni Magnussen Institute for Icelandic Studies (2016). ISLEX: An Islandic – Scandinavian Multilingual Dictionary. http://www.arnastofnun.is/page/islex_en (Accessed: 12 March 2016).

The GeoHumanities Special Interest Group: Fostering and facilitating the geospatial turn

Katherine Hart Weimer
kathy.weimer@rice.edu
Rice University, United States of America

Karl Grossner
karlg@stanford.edu
Stanford University, United States of America

Introduction

The GeoHumanities Special Interest Group (SIG) was formed in 2013, and in two and a half years, over 500 individuals have subscribed to the group's email and social media feeds. However, there remains considerable untapped potential for SIG activities informing and enabling humanities scholars' investigations of geospatial research methods and theory. This poster will publicize the SIG's mission and activities, and gather input from conference attendees on possible future endeavors.

Formation

The 2013 Digital Humanities conference program reflected an increasing interest in questions of place and space by scholars from various disciplines. Recognizing the need for cross-disciplinary collaborations and sharing of experience, the SIG co-founders responded to ADHO's call for proposals during that year's SIG slam. Our intent was to utilize ADHO's SIG structure to unite and support the disparate groups of scholars conducting research with a spatial or geographic context. In naming the SIG, we have emphasized geography's traditional attention to not only spatial, but spatial-temporal and 'placial' perspectives. Soon after DH2013, the GeoHumanities SIG proposal was approved.

Goals

The SIG's stated goals are

> "... to create a venue for pooling knowledge and best practices for relevant existing digital tools and methods, to foster the collaborative development of shared resources and new tools and extensions to geospatial software, and to keep humanist scholars at large informed about the possibilities and inherent pitfalls in their use."

Activities

The SIG uses various communication channels, including a web site (geohumanities.org), an email-list and twitter feed. Subscribers now number over 500, primarily from the US and Europe, and the group has had a few notable accomplishments during that time. SIG members have contributed to the expansion of the DiRT Directory by augmenting the TaDiRAH DH taxonomy and are contributing annotations to DiRT listings that are relevant to geospatial research, on an ongoing basis. Those listings are fed automatically to the SIG site in a feature called *GeoDiRT*. The web site also features a list of over 150 Humanities GIS projects from around the world. That resource was originally developed by one of our members several years ago, and transferred to our ADHO-hosted site to ensure its longevity and to facilitate contributions. The SIG website includes blog postings, announcements and a twitter feed.

In each of the past two years the GeoHumanities SIG has held pre-conference events. At Lausanne in 2014, a well-attended day-long meeting focused on "Place and Period in an Emerging Global Gazetteer," and at Sydney in 2015, the SIG promoted a workshop hosted by several of its members on the topic of peer review in GeoHumanities. In 2016, the SIG endorsed and promoted a member organized workshop, "A Place for Places: Current Trends and Challenges in the Development and Use of Geo-historical Gazetteers." Each of these sessions provided a venue to share expertise and shape research agendas.

Future

The GeoHumanities SIG's future initiatives may include:

- Enabling the enhancement of Humanities GIS project listings by their investigators.
- Improving browse and search capability for the Humanities GIS project list.
- Improving integration of DiRT and GeoDiRT.
- Coordinating the collaborative development of an online primer on geospatial analytic and mapping tools, data resources, and best practices.
- Engaging SIG members to enhance GeoDiRT by reviewing tool listings.
- Other ideas brought forward by members.

The DH conference poster venue is a great opportunity to increase awareness of the GeoHumanities SIG and its activities, expand the member base and gather input on current and future directions.

DH Bridge: Teaching Computational Thinking in the Humanities

Jeri Wieringa
jwiering@gmu.edu
George Mason University

Celeste Sharpe
csharpe2@masonlive.gmu.edu
University of Pennsylvania; George Mason University

This poster will introduce and demonstrate DH Bridge (http://dhbridge.org/), a pedagogical model and curriculum for one to two day workshops that provides intensive training in computational thinking in humanities contexts. There are an ever increasing number of opportunities for training in the digital humanities, from informal workshops held at THATCamps (http://thatcamp.org/) to multiple week training programs, such as the NEH Institutes for Advanced Topics (http://www.neh.gov/grants/apply-neh-funded-seminar-institute-or-workshop), Humanities Intensive Learning and Teaching (HILT) (http://www.dhtraining.org/hilt/), and the Digital Humanities Summer Institute (DHSI) (http://www.dhsi.org/). While these workshops provide excellent opportunities to gain exposure to new ideas or to dive deep into a particular subject area, they are also dependent upon the expertise of the particular instructors and, depending on the length of the workshop or institute, require substantial commitments of time and money.

Building on the examples of Rails Bridge (http://www.railsbridge.org/) and Rails Girls (http://railsgirls.com/), which focus on lowering the barriers to entry for underrepresented persons in code, DH Bridge offers an additional model for training in the digital humanities. Rather than relying on a single instructor, DH Bridge relies on a group of coaches to guide participants through a shared tutorial, troubleshoot technical problems, and provide additional contextual information for participants. Throughout the day of the workshop, technical work is interwoven with group activities aimed at helping participants develop useful mental models of computational processes and articulate research questions that would benefit from computational analysis. We have found that this combination of community supported learning, technical work, and group exercises creates a supportive environment where technical learning can take place.

DH Bridge started as a locally hosted Rails Girls workshop for humanities scholars (http://railsgirls.com/digitalhumanities_fairfax) held in early September 2013. The event was well-received and successful, highlighting for us the strength of the model of the short workshop with a central curriculum and small group coaching. It was clear, however, from the feedback we received that the Rails Girls curriculum was too focused on gaining technical skills for employment, as participants struggled to connect the skills they learned to their intellectual goals as humanities scholars. Thanks to an incubation grant from the Association for Computers and the Humanities (http://ach.org/2014/07/09/ach_microgrants_winners_2014/), we were able to develop a curriculum focused not on the mechanics of code, per se, but on the processes of computational thinking. This curriculum was used in a second workshop, held on November 1, 2014.

What is unique about this pedagogical approach in the current landscape of digital humanities training is that the workshops run from a central and open curriculum that can be adjusted and expanded for a particular community. Because the emphasis is on patterns of thinking, rather than use of particular tools, the workshop curriculum is flexible. Whether using the included text-mining tutorial, or adapting tutorials from outside sources such as The Programming Historian (http://programminghistorian.org/), workshop organizers can customize the day for the interests of their community. In addition, the central curriculum, together with the local focus and short timeframes of the events, helps to keep the costs of the workshops low for both participants and organizers. And, most importantly, the success of the curriculum is derived from the community—the participants and the coaches work together to solve problems, both at the level of the code and at the theoretical level of scholarly investigation through code.

We've taken this approach to help expand the community of scholars leveraging computational approaches in their research. Cultural and institutional barriers to learning computational methods are well documented, and organizational responses like FemTechNet (http://femtechnet.org/) and GO::DH (http://www.globaloutlookdh.org/) are doing valuable work in challenging the systemic barriers. DH Bridge aligns with this work, and seeks to provide a local means for groups of people underrepresented in the digital humanities to learn and engage with the concepts and skills of programming in a way that is meaningful to their interests.

This poster will address our lessons learned from running the two iterations of the workshop, as well as our plans for future workshops. We will discuss the model of the one to two day workshop as a way of lowering the costs for participants and organizers; our theoretical and pedagogical choices in developing the curriculum; the challenges we faced and some lessons learned; and the feedback we have received from participants. We invite and look forward to engagement with and collaboration on the curriculum and the model of DH Bridge.

Browsing, Sharing, Learning and Reviewing the Haine du théâtre Corpus through Insightful Island

Bin Yang
yangb86@hotmail.com
Université Pierre et Marie Curie-Sorbonne Universités; Laboratoire d'informatique de Paris 6

Chiara Mainardi
chiara85.mc@gmail.com
Université Paris-Sorbonne-Sorbonne Universités; Labex OBVIL

Jean-Gabriel Ganascia
Jean-Gabriel.Ganascia@lip6.fr
Université Pierre et Marie Curie-Sorbonne Universités; Laboratoire d'informatique de Paris 6; Labex OBVIL

Researchers of DH need visualizations to help them analyze and share their research results. Meanwhile, the users of the corpus are waiting for a tool (e.g. knowledge Map) to help them with learning and information seeking, and help them with knowledge understanding and memorizing tasks. The insightful island is one of the most promising tool for the perception of the corpus and its memorization.

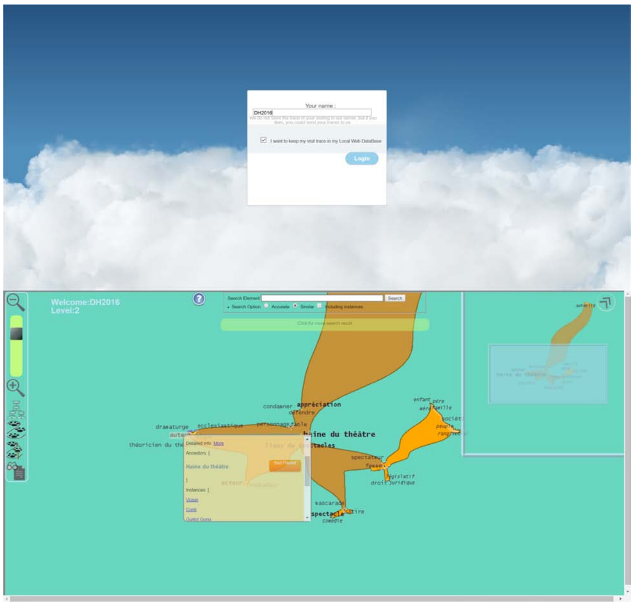
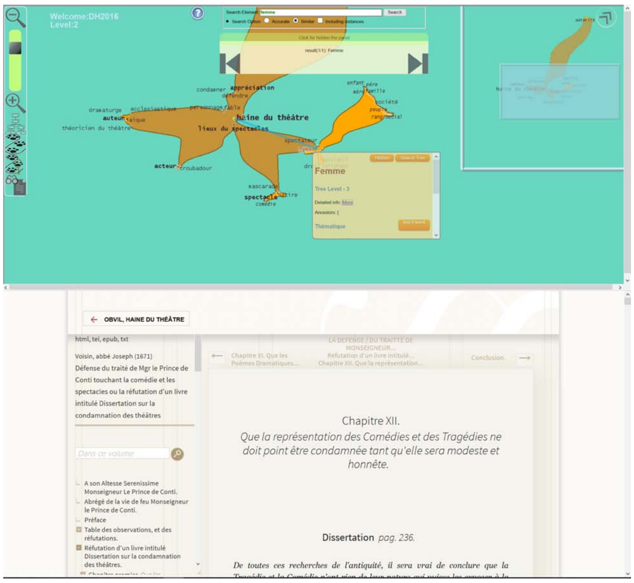


Figure 1. Visualize the HdT corpus as an insightful Island

Cartographic Visualizations (Gansner et al., 2013), have a lot of advantages: they propose to map contents onto two-dimensional knowledge maps that use cartographic metaphors and geographical analogies. These help users to understand the knowledge and allow them to show their own understanding on the visualization, then share it with other scholars. With knowledge maps, people from different domains can collaborate with each other, thus leading to additional benefits. For example, the famous Torrance's experiments (Torrance, 1970) show that working in pairs facilitates creativity.



← OBVIL, HAINE DU THÉÂTRE

html, tei, epub, txt

Voisin, abbé Joseph (1671)
Défense du traité de Mgr le Prince de Conti touchant la comédie et les spectacles ou la réfutation d'un livre intitulé Dissertation sur la condamnation des théâtres

Dans ce volume

- A son Altesse Serenissime Monseigneur Le Prince de Conti.
- Abrégé de la vie de feu Monseigneur le Prince de Conti.
- Préface
- Table des observations, et des réfutations.
- Réfutation d'un livre intitulé Dissertation sur la condamnation des théâtres.

← Chapitre XI. Que les Poëmes Dramatiques... LA DEFENSE / DU TRAITTE DE MONSEIGNEUR... Réfutation d'un livre intitulé... Chapitre XII. Que la représentation... Conclusion. →

Chapitre XII.
Que la représentation des Comédies et des Tragédies ne doit point être condamnée tant qu'elle sera modeste et honnête.

Dissertation *pag. 236.*

De toutes ces recherches de l'antiquité, il sera vrai de conclure que la

Figure 2. Access, detail information and search with the HdT island

The Project *Haine du théâtre* (HdT) aims at analyzing theatre debates in Europe by using scientific approaches and critical editions of polemical texts. The reflections of the HdT team are mainly focused around the discovery of the circumstances and the arguments used in theatre controversies all across Europe, from the 16th century up to the 19th century.

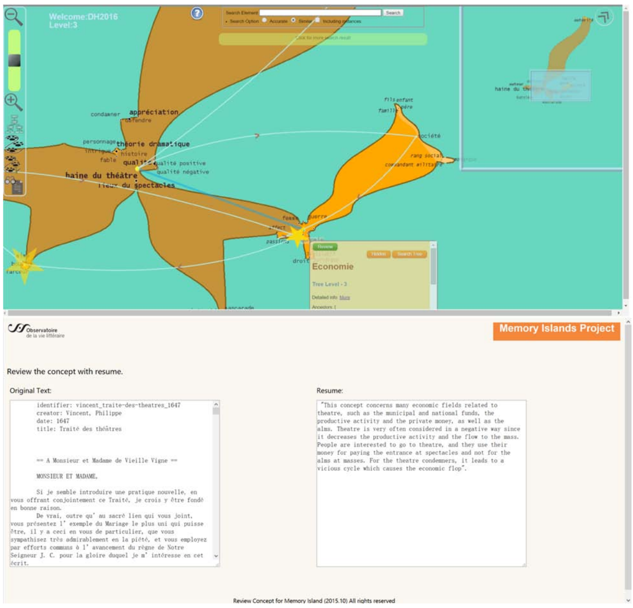
In this poster, we show how to use Memory Island (MI) technique (Yang, 2015) to help us creating an insightful island for the HdT corpus. MI is inspired by the "loci" method of the Art of Memory technique, and it consists of associating each entity of knowledge to a designated area on a created virtual island. We generated a 2D knowledge island (Figure 1) of an ontology built to access the HdT corpus by employing the MI technique.

This visual representation allows users to navigate through the corpus, based on the insights of experts. It uses distance measures, based on the perception of the literary field expressed by scholars specialized of the considered period. This is achieved in this same way that experts from the "Knowledge Cartography" field (Okada et al., 2008) manually craft their knowledge maps.

Users can circulate through the corpus and discover interesting documents that appear to be semantically close to the one they are viewing. The overall organization being based on the ontology skeleton (i.e. a hierarchical structure), the spatial proximity corresponds to a semantic proximity between documents that could arise unexpectedly to the eyes of the viewer, which should stimulate his/her curiosity. In addition, the users can get more information about the concepts by simply clicking their labels on the map: detailed information windows appear on demand. Moreover, the users can study the online corpus, together with the labels of instances and the supplementary information by clicking the chosen items in this window (Figure 2).

Then we designed two different actions: "Visit" and "Study". "Visit" is the action performed by the user that allows her to see a concept's detailed information; "study" is the action allowing users to examine concepts and learn from it at a deeper level. We developed techniques for helping the users to learn and memorize this knowledge based on these map representations. For example, the users can share visiting trace with others, which allows to collaboratively study the corpus (Figure 3). This is essential in the perspective of a participatory activity in DH. User can visually display her visiting trace and allow others to re-visit her trace.

The studied concepts of one user are also visually represented (Figure 3) by the novel review function which simulates our mind's memorization and learning mechanisms. The frequency of learning is also visually represented, using degrees of transparency (least transparent means most studied), in a technique analogous to that of the "heat-maps". If the user wants to review the studied concepts, it will show her a summary about the concept, prepared by the expert, also allowing the comparison with the original text in the corpus.



Review the concept with resume.

Original Text:

```
        identifier: vincent_traite-des-theatres_1647
        creator: Vincent, Philippe
        date: 1647
        title: Traité des théâtres



        == A Monsieur et Madame de Vieille Vigne ==

        MONSIEUR ET MADAME,

        Si je semble introduire une pratique nouvelle, en
vous offrant conjointement ce Traité, je crois y être fondé
en bonne raison.
        De vrai, outre qu' au sacré lien qui vous joint,
vous présentez l' exemple du Mariage le plus uni qui puisse
être, il y a ceci en vous de particulier, que vous
sympathisez très admirablement en la piété, et vous employez
par efforts communs à l' avancement du règne de Notre
Seigneur J. C. pour la gloire duquel je m' intéresse en cet
écrit.
```

Resume:

```
"This concept concerns many economic fields related to
theatre, such as the municipal and national funds, the
productive activity and the private money, as well as the
alms. Theatre is very often considered in a negative way since
it decreases the productive activity and the flow to the mass.
People are interested to go to theatre, and they use their
money for paying the entrance at spectacles and not for the
alms at masses. For the theatre condemners, it leads to a
vicious cycle which causes the economic flop".
```

Review Concept for Memory Island (2015.10) All rights reserved

Figure 3. Visualize the visiting trace and review the studied concepts (highlighting with the five-point stars). In this case, the user has studied "society", "quality" and "thematics" concepts, alighting on the concept "economy" and reading its summary

In the future, we envisage integrating Natural Language Processing and Artificial Intelligence techniques, such as Name Entities Recognition to rendering our technique fully automatic; this in turn will enable all scholars in DH to develop their own Memory Islands for sharing their knowledge more easily.

Bibliography

Gansner, E., Hu, Y., and Kobourov, S. (2013). *Handbook of Human Centric Visualization*. Springer.

Okada, A., Shum, S. B., and Sherborne, T. (2008). *Knowledge Cartography: Software Tools and Mapping Techniques (Advanced Information and Knowledge Processing)*. Springer.

Torrance, E. P. (1970). Dyadic interaction as a facilitator of gifted performance. *Gifted Child Quarterly*, **14**(3): 139–43.

Yang, B. (2015). *Memory Island: Visualizing Hierarchical Knowledge as Insightful Islands*. Ph.D thesis, University Pierre and Marie Curie.

Mapping Kipnis

Sarah Ellen Zarrow
sarah.zarrow@gmail.com
New York University, United States of America

Hanna Kipnis King
Hannashimona@gmail.com
Harvard University Libraries

What did everyday life look like for a Polish Jew in Warsaw around the time of the First World War? We teach music history, musicology, Jewish history, and Slavic studies. Our students come to us with various pre-conceptions: either that Jews and Poles (conceived of as mutually exclusive categories) lived in open hostility to each other, or that they lived entirely separate from each other, their paths only crossing out of necessity. We want to create, together with our students, a digital tool that will challenge this pre-conception, and their broader ideas about history. We know that preconceptions often interfere with classroom instruction, and often remain much more salient than new knowledge (see, for example, Sam Wineburg, *Historical Thinking and Other Unnatural Acts*, 2001). We believe that hands-on digital tools may do more to interfere with preconceptions than traditional methodologies.

The work of Menachem Kipnis (1878-1942) belies those assumptions. For 16 years, between 1902 and 1918, he performed with the Warsaw Opera; he also sang at the a synagogue in downtown Warsaw. He wrote about these experiences in a variety of newspapers, including Poland's widest circulation Yiddish-language daily, *Haynt*. His accounts are full of rich details that mix the personal and the geographic—how he could run from Tłomackie Street to the Opera and arrive, out of breath, but with enough time to change, for example. Other archival records in which he appears demonstrate the difficulties that Jews faced in joining non-Jewish Polish society—Kipnis's obviously-Jewish first name was not always given in Opera programs, or is Slavicized.

We came to create *Mapping Kipnis* as a challenge to see what an interactive mapping project, one designed by historians and Jewish studies scholars, could do for students' (and our own) understandings of Kipnis's world. Using Omeka, we present Kipnis's Warsaw, with locations enhanced by his words and photographs. We plan a second phase of the project to look at Kipnis's tour through Eastern Europe; his writing from those years offers a warm and colorful picture of Jewish life before the Holocaust. This phase will help students form an understanding of the political and cultural geography of Europe during the period.

Virtually none of our students have been to Warsaw, and have only the vaguest understanding of what it looks like, and even less of an understanding of pre-World War Two Warsaw. For these students, we believe that using interactive tools can provide a layer of understanding about the social geography of Jewish life in Poland that a purely textual account cannot.

Mapping Kipnis is incomplete and we envision students being able to contribute to it, and to add adjacent projects to it as well. Thus, it is both an instructionally demonstrative tool, and an opportunity for our students to familiarize themselves with digital mapping and with the potential digital life of archival material. We are delighted to bring *Mapping Kipnis* to Krakow in 2016, for demonstration and feedback. We hope it sparks as many conversations among participants as it has among the three of us.

Towards a Cross-generation Social Network for Jewish Sages

Maayan Zhitomirsky-Geffet
Maayan.Zhitomirsky-Geffet@biu.ac.il
Bar Ilan University, Israel

Gila Prebor
gila.prebor@biu.ac.il
Bar Ilan University, Israel

AmichaI Feigenboim
fgnbmcs@computing-services.co.il
Bar Ilan University, Israel

Introduction

Construction and analysis of social networks for historical figures has lately become a popular approach in History and Prosopography (Keats-Rohan, 2007), Sociology (Wetherell, 1998; Roldán Vera and Schupp, 2006) and digital humanities (Rochat, 2015; Yamada, 2015). This approach is especially beneficial for providing a global view and automatic mathematical and statistical analysis for large historical corpora (Rossi et al., 2013), for which researchers are unable to gain much knowledge by even an exhaustive manual exploration.

Jewish Biblical and Rabbinic literature is a great source of ancient wisdom and cultural heritage. It includes a large amount of people such as prophets, political and religious leaders, sages and other historical figures. Amazingly, although these people were spread over the world and through different time periods, they were united and connected by the same text - the Bible. Therefore, the aim of this research is to propose and implement a methodology for construction of a cross-generation social network for Jewish sages to explore their inter-relationships on a large scale, using modern computerized tools for text analysis and graph mining (Rossi et al., 2013; Yamada, 2015).

The proposed methodology

At the first stage we define the corpus of the study and a reliable digital resource for this corpus. We work with the text of Mishna (2nd century CE) and Talmud (4th-5th century CE). Next, the following information is retrieved from existing traditional research sources, such as encyclopedia of Jewish sages (as most of these sources have not been digitized, the person-related data is extracted manually and stored in the digital form):

1. A list of sages for the selected corpora. One of the biggest challenges with sages' names is their ambiguity and a large number of namesakes (Rochat, 2015; Keats-Rohan, 2007). To tackle this problem we add identifying discriminative features to each name (e.g. father's name or place of birth).

2. A list of basic relationships between sages, e.g. family relationships, teacher-student, time period, place, possessing a similar political/social/professional role, studying in the same institution, participation in the same event.

Finally, the above basic relationship list can be further extended with text-based relationships, such as sages who cite each other, disagree, or comment on the same section of the biblical text. This is achieved by automatically learning lexical patterns in which pairs of sages co-occur in texts and using them to extract the corresponding relations.

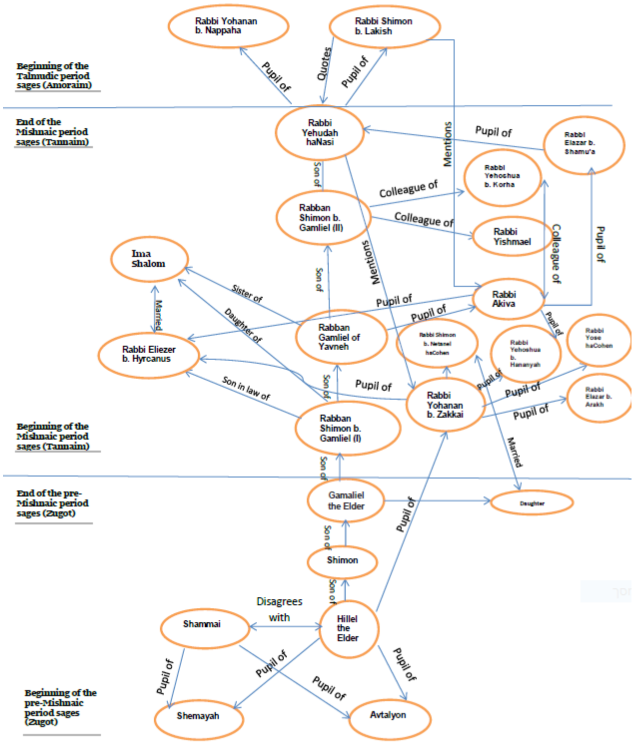


Figure 1: A fragment of the cross-generation social network for Jewish sages

All the extracted data from multiple independent resources are digitized and integrated in the single database and can be queried and visualized by the common tools (e.g. Gephi (Bastian et al., 2009)). Figure 1 illustrates a fragment of the proposed type of the cross-generation social network for Jewish sages. Complex queries can be further answered by mining the network, e.g. whether a given pair of sages are related and how? What are all the various direct and indirect relationships of a given sage? Whether the same text segment cites sages from different time periods (meaning that it has been edited at a later period)? At the global level the social network helps identify the central figures/communities of the sages in different places, times, schools, dynasties, philosophical approaches, text segments, and citations according to the number of network relationships and their density, centrality and coreness (Rochat, 2015; Keats-Rohan, 2007). The historical data in the built network becomes accessible to researchers from the humanities and will take their research capabilities to the next level.

Bibliography

Bastian, M., S. Heymann, and M. Jacomy. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Proceedings of the AAAI Conference on Weblogs and Social Media*, Eytan Adar et al. (Eds.), Menlo Park: AAAI Press, pp. 361-62.

Barber, Michael J. (2007). Modularity and Community Detection in Bipartite Networks. *Physical Review E,* **76**: 1-9.

Keats-Rohan, K.S.B. (2007). *Prosopography Approaches and Applications: A Handbook*. Oxford: Linacre College Unit for Prosopographical Research.

Rochat. Y. (2015). Character network analysis of Émile Zola's Les Rougon-Macquart. In *Proceedings of DH 2015*, Sydney.

Roldán Vera, E. and T. Schupp. (2006). Network analysis in comparative social sciences. *Comparative Education* **42**(3).

Rossi, F., Villa-Vialaneix, N. and Hautefeuille, F. (2013). Exploration of a Large Database of French Notarial Acts with Social Network Methods. *Digital Medievalist,* 9.

Wetherell, C. (1998). Historical Social Network Analysis. *International Review of Social History*, **43**: 125-44.

Yamada, T. (2015). Detection of People Relationship Using Topic Model from Diaries in Medieval Period of Japan. In *Proceedings of DH 2015*, Sydney.

Pre-conference workshops

RDA/ADHO Workshop: Evaluating Research Data Infrastructure Components and Engaging in their Development

Bridget May Almas
bridget.almas@tufts.edu
Tufts University, United States of America

Kim Fortun
fortuk@rpi.edu
Rensselear Polytechnic Institute, United States of America

Natalie Harrower
n.harrower@ria.ie
Digital Repository of Ireland, Ireland

Eveline Wandl-Vogt
Eveline.Wandl-Vogt@oeaw.ac.at
Austrian Academy of Sciences, Austria

Summary

The purpose of this workshop is to conduct a meaningful examination of the data fabric and infrastructure components being defined by the Research Data Alliance (RDA), to test their relevance and applicability to the needs of the digital humanities community, and to discuss opportunities for humanities engagement in further standards development.

RDA is an international initiative to facilitate the development of effective data practices, standards and infrastructure in particular research domains, and across domains. It aims to enhance capacity to archive, preserve, analyze and share data, and for collaboration both within and across research communities. The humanities have an important presence in RDA, and can benefit from the opportunities RDA provides to learn across research communities working to develop digital infrastructure. RDA also brings together diverse types of technical expertise, which is organized to put forward (best practice) "adoption products." Some of these products are starting to be taken up in the humanities, such as the Practical Policy Recommendations mentioned below, and there is significant potential for further collaborative work between RDA and digital humanities developers in the future.

Much of the infrastructure needed to support data sharing is of great relevance to Digital Humanities projects, where we find ourselves too often developing and reinventing ad-hoc solutions for data management, draining resources that could be put to better use focusing on the domain-specific nature of our problems and driving new research. It's easy, especially when time and resources are constrained, to get locked into thinking that our problems are unique and that we need to design custom solutions, but when we examine the problem from other perspectives, the abstractions begin to rise to the surface. But in order to take advantage of the solutions as they are built, we must be part of the discussion about the requirements, push for our use cases to be considered in their design, and take part in testing, implementing and sustaining the solutions.

This will be a full day workshop in the format of a hands-on round-table and open discussion. Participants will be asked to come prepared to discuss details of their particular use cases, as well as solutions and needs relevant to two of the initial outputs of the RDA: the Persistent Identifier (PID) Types (Weigel, et. al., 2015) and Data Types Registry (DTR) (Lannom, et. al., 2015). In advance of the workshop, organizers will provide summaries of these outputs and detailed examples of their analysis for use cases in humanities and other relevant domains.

This workshop is a complement to the panel by Dr. Natalie Harrower et. al. entitled "Digital data sharing: the opportunities and challenges of opening research". The panel presents particular challenges in humanities research data management, and aims to generate a discussion around the uniqueness and challenges inherent in humanities research data. This workshop, on the other hand, is a hands-on effort to work with real humanities data use-cases, provided by participants, to understand how to best shape RDA outputs to enable better data sharing and management in the humanities.

Format of the Workshop

In the first two hours of the workshop, organizers will present a summary of humanities activities in RDA thus far, and describe current calls for participation by RDA working and interest groups.

These calls address an array of topics important in the digital humanities: Institutional Review Boards (IRB), access solutions and metadata standards that support data sharing; the need for institutional repositories for both live and archived digital projects; the need for a "data net" to connect globally distributed repositories, enabling discovery and access; the need for cultural and organizational changes in the humanities to research data sharing and open scholarship.

Specific RDA activities covered will include:

- The Digital Practices in History and Ethnography Interest Group (DPHE-IG) (https://rd-alliance.org/node/508), chaired by anthropologists Mike Fortun and Kim Fortun at Rensselaer, and Jason Jackson, Director of the Mathers Museum of World Cultures at the University of Indiana.
- The successful adoption by the Platform for Experimental and Collaborative Ethnography (PECE) (http://worldpece.org) of the RDA Practical Policies

Recommendations, a specification for best practices for data management.

- Two nascent RDA Working Groups: the Research Data Collections WG (https://rd-alliance.org/groups/pid-collections-wg.html) and the WG on Empirical Humanities Metadata.

We will wrap up the first half of the morning session with a group effort to identify the range of solutions and support needed for research data management and sharing in the digital humanities in coming years, and potential opportunities for RDA collaboration. The list generated will be taken back to the RDA community for their consideration and feedback.

The second half of the morning will be devoted to an in-depth presentation of the RDA PID Types and DTR outputs, including a demonstration of their implementation.

After lunch, each participant will be invited to present their use case/requirements and engage with workshop organizers and participants in a discussion of the relevance and gaps and cost/benefit of adopting the solutions being proposed by RDA. Prior to the workshop the organizers will issue a short survey for participants to answer directed questions about their requirements as well as a template for more detailed descriptions of their use cases. Specific focus will be on the PID Types and DTR solutions, but other relevant outputs or in-progress efforts may be considered as well. Workshop organizers will take notes and produce a summary report following the workshop to share with the RDA community for their consideration and feedback.

Target Audience

Members of the ADHO community who are interested in collaborating with a global multi-disciplinary community to define, develop, test and adopt infrastructure for supporting the management, preservation and sharing of humanities research data. Participants should have some experience with digital humanities projects for which general solutions for working with Persistent Identifiers and machine actionable Data Types are relevant.

Workshop Leaders

Bridget Almas is a Senior Software Developer for the Perseus Digital Library at Tufts University (http://www.perseus.tufts.edu) and co-PI of the Perseids Project (http://www.perseids.org), a collaborative online environment for creating and publishing datasets consisting of transcriptions, translations, linguistic annotations and commentaries of and on ancient source documents. She is a co-chair of the RDA Research Data Provenance Interest group, and a former member of the RDA Technical Advisory board.

Kim Fortun is a cultural anthropologist and Professor of Science & Technology Studies at Rensselaer Polytechnic Institute. From 2005-2010, Fortun co-edited the *Journal of Cultural Anthropology* (http://www.culanth.org/), as it was developing its original digital infrastructure. Fortun has played a lead role in the development of the Platform for Experimental and Collaborative Ethnography (PECE), an open source/access online work space for anthropological and historical research. Fortun co-chairs the RDA DPHE Interest Group.

Dr. Natalie Harrower is the Director of the Digital Repository of Ireland (DRI) (http://www.dri.ie/) - a publicly accessible online repository for the long-term preservation, sharing and reuse of humanities and social science data. In addition to building an open-source trusted digital repository to international standards, the DRI also influences policy and publishes reports and guidelines on data archiving, metadata standards, preservation infrastructures, Linked Data, digital preservation challenges unique to cultural data, and digital humanities. DRI is piloting a research data management project with diverse digital arts and humanities data, and their current flagship project is the award-winning digital cultural heritage site *Inspiring Ireland* (inspiring-ireland.ie).

Eveline Wandl-Vogt, research manager at the Austrian Academy of Sciences, Austrian Centre for digital Humanities and VCC1 Co-Chair eInfrastructures, COST ENeL; expert in several national and international committees – mainly focussing on standardisation, interoperability, Social Innovation and Open Science. Recently, she mainly focuses on supporting transformation processes from the Humanities towards Interdisciplinary Humanities and applied Humanities in the framework of Open Science, Citizen Science and Open Innovation.

Bibliography

Lannom L., Broeder, D. and Manepalli, G. (2015). *Data Type Registries Working Group Output*, https://rd-alliance.org.

Weigel T., DiLauro, T. and Zastrow, T. (2015). *PID Information Types WG Final Deliverable*, https://rd-alliance.org.

A Place for Places: Current Trends and Challenges in the Development and Use of Geo-Historical Gazetteers

Carmen Brando
carmen.brando@ign.fr
Institut National de l'information géographique et forestière (IGN), France

Francesca Frontini
francesca.frontini@ilc.cnr.it
Istituto di Linguistica Computazionale "A.Zampolli", Italy

The implementation of geo-historical gazetteers increasingly depends upon the development of Natural Language Processing (NLP) and Corpus Linguistics as well as geographical analysis in disciplines such as History, Archaeology and Literary Studies. The application of these methods usually relies on the appropriate modelling of databases for performing the semantic enrichment of documents including geoparsing tasks. At the same time, even when performing a manual enrichment and referencing of place mentions in texts or in library or museum catalogues (e.g. when applying the CIDOC CRM model and its spatio-temporal extension), an adequate source of external information is crucial.

Today, geo-historical data are more and more often published following the Linked Data (LD) principles: i.e. using URIs and data format standards (RDF) and linking to other data sets to enable information discovery. Moreover, an implicit driving principle of LD, widespread in the Semantic Web community, is the reuse of vocabularies and ontologies already defined by others to avoid duplication. Keeping track of provenance is crucial. Pleiades is one of the best examples these days, but other generalistic sources such as DBpedia, Wikidata or GeoNames also provide interesting - albeit partial – geo-historical information and have proved to be useful in Digital Humanities (DH) projects. Linking texts to external sources using URIs enables the retrieval of additional information about the referenced places. Once this has been achieved, the information in the sources can be easily used to produce different views and aggregated analysis of corpora: i.e. visualizations (Jessop, 2008); this in turn is meant to help scholars to capture place perceptions and to analyse spatio-temporal phenomena described in corpora.

The choice of geo-historical datasets which are used as gazetteers depends on the domain of the texts under consideration. Pleiades is specifically suited to places in Mediterranean Ancient History texts. However, tasks such as referencing places from historical periods other than Antiquity, or identifying geographically vague or imaginary places in literary texts, if ever possible, might need the development of a different methodological approach, which would include the construction of conceptual mapping models and the creation of a completely different kind of gazetteer. In any case, the choice will have an important influence on the results of such visualizations as well as on the pertinence of the interpretation. Existing gazetteers vary widely in how they abstract the world. Important aspects – such as scale, the representation of time (and change over time), complex geometries, uncertainty and vagueness as to location and/or date, multiple points-of-view, representation of hierarchies of political-administrative units, their boundaries and their change over time, alternative names (rejected and standard forms, vernacular and multilingual), representation of fantastic places – are modelled in different ways, or are missing altogether. This limits their applicability in the Humanities. Moreover, interlinking between corresponding entities in different gazetteers is often lacking, although progress has been made in this regard, through community initiatives or by using GeoNames, or Wikidata as backbones (Simon et al., 2015). Finally, the ontologies used to link toponyms in texts to spatial references need to be further developed, especially when it comes to deal with fuzziness and uncertainty in mentions (Reuschel and Hurni, 2011).

Clearly, new models should conform to LD principles, and they should privilege the reuse of existing and consolidated ontologies, vocabularies and datasets whenever possible. Long term preservation and maintenance are also crucial problems in this sense because texts enriched with references to sources that have become obsolete or unavailable may have results that are unusable for the task for which they were tagged (Janowicz et al., 2012). In this sense, specialisation of efforts on the one hand (for pooling efforts) and coordination on the other, are crucial for such projects. Finally, geo-historical projects should also promote harmonization of their data with standards and practices of the broader DH community, and of the current research trends, in particular for what concerns the interoperability of resources within the framework of larger research infrastructures such as CLARIN or DARIAH.

In the proposed full-day workshop, we will focus on geo-historical gazetteers, and we will discuss their limits in supporting the needs of the Spatial Humanities (SH) community. The proposed workshop will be composed of nine presentations (abstracts are listed below), each of which concerns the production of geo-historical gazetteers as LD as well as the annotation, the recognition and the geoparsing of place names referenced in texts, library and museum catalogs, digitized maps, etc.

Christopher Donaldson (University of Birmingham), Extracting and visualizing the geographies in historical travel writing: This presentation will introduce a procedure for the automated extraction and resolution of geographical information from a corpus of historical writings about the English Lake District. The research on which the presentation is based is using the spatial analysis of geo-historical LD sets to achieve a more comprehensive and refined understanding of how the landscape of the Lake District was perceived, represented, and experienced in the past.

Karl Grossner (Stanford University), Joining Place and Period in Historical Gazetteers: Places referred to in historical documents and gazetteers have temporal as well as spatial extents. Likewise, historical periods have spatial extents. However existing data models and format standards and the mapping and timeline software that use them do not reflect this. I will discuss recent work on Topotime, an extension to the GeoJSON format adding temporal expressions, and allowing for some types of uncertainty encountered in historical data.

Katherine Hart Weimer (Rice University): A wealth of geographic information is included in library catalogs, with existing structures for name disambiguation, cross-referencing and inclusion of geographic coordinates. Recently, efforts are underway in libraries to convert this data into LD allowing for cross platform applications. The presentation describes an experiment in this sense.

Maurizio Lana (Università del Piemonte Orientale): Annotation of place mentions in Latin Literature. The annotation pipeline uses parsing+NER but later mentions are manually checked and referenced to external gazetteers such as Pleiades. The novelty of the project is the GeoLat GO! ontology that allows for a more complex annotation.

Bruno Martins (University of Lisbon): NLP and IR methods for handling geospatial information in textual documents. In my talk, I will present a brief survey of techniques for handling geospatial information within textual documents, including work at our team in the University of Lisbon, and other methods proposed within the Computational Linguistics and IR communities. I will discuss methods to address the problems of (i) document geocoding, (ii) toponym resolution, and (iii) selecting geographically relevant key-phrases. Applications within the broad field of DH, and SH in particular, will also be outlined.

Patricia Murrieta-Flores (University of Chester): So far, research in the SH has been mainly concerned with geographically precise information or what could be considered as 'real' places in historical and literary sources. Nevertheless, non-locational places play an important role in narratives of all sorts of sources from the fantastic, to geographically vague travel accounts. This is an important limitation in the analysis of place in the DH. Using Medieval Romances as an example, this presentation will discuss the challenges posed by literary narratives of place in terms not only of disambiguation, but also reference to fantastic and non-locational places.

Michael Page (Emory University): Atlanta Explorer: Historical Geocoding & the City: Atlanta Explorer focuses on building datasets and geospatial tools to explore the history of the city. Completed is a geodatabase and geocoder for circa 1930 and the pilot 3D virtual environment. The next phase includes producing geocoders for the remaining years (1868–1930) and therefore strategies and methods for developing historical geocoding datasets and tools for place discovery will be discussed. Our goal is to also share the underlying data with the community CityGML as how we would likely share and archive the model.

Rainer Simon (Austrian Institute of Technology), Pelagios project: an international community initiative concerned with the development of Linked Open Data methods, tools and services to better interconnect geo-historical datasets. In its most recent project phase ("Pelagios 3 - Early Geospatial Documents"), Pelagios has developed Recogito, a semi-automatic geo-annotation tool; Peripleo, a geotemporal search engine. Furthermore, Pelagios has annotated more than 300 historical sources from different cartographic traditions, collecting more than 120,000 place references in literary texts and early maps.

Humphrey Southall (University of Portsmouth): Engaging the wider public with historical gazetteers. Gazetteers are a powerful tool for humanities researchers, but are also of great fascination and utility for the general public. That interest enables academic projects to achieve wider "impact", enables popular web sites to be sustained by advertising income, and enables expansion through crowd-sourcing. This presentation covers three experiences: the established Vision of Britain site; PastPlace, our new LD gazetteer which uses Wikidata as a spine to which we are adding historical toponyms; and GB1900, a crowd-sourced gazetteer building project developed in collaboration with National Libraries in Great Britain.

The proposed workshop targets an audience of scholars, data designers, and software developers, and will also comprise a speed presenting session for participants, topic-based breakout discussions between experts and attendees, a panel to highlight research priorities and summarize the main contributions of the workshop and research directions.

Bibliography

Berman, M., Mostern, R. and Southall, H. (2016). *Placing Names: Enriching and Integrating Gazetteers*. Bloomington, IN: Indiana University Press.

Elliott, T. and Gillies, S. (2009). Digital Geography and Classics. *Digital Humanities Quarterly*, **3**(1).

Evans, C. and Jasnow, B. (2014). Mapping Homer's Catalogue of Ships. *Literary and Linguistic Computing,* **29**(3): 317–25. doi:10.1093/llc/fqu031.

Gregory, I., Donaldson, C., Murrieta-Flores, P. and Rayson P. (2015). Geoparsing, GIS and textual analysis: Current developments in Spatial Humanities research. *International Journal of Humanities and Arts Computing*, **9**: 1–14. See: DOI: 10.3366/ijhac.2015.0135

Grossner, K.,Janowicz, K. and Keßler, C. (2014). Place, Period, and Setting for Linked Data Gazetteers. In Berman, M., Mostern, R. and Southall, H. (Eds.), *Placing Names: Enrich-*

ing and Integrating Gazetteers. Bloomington, IN: Indiana University Press. http://geog.ucsb.edu/~jano/GrossnerJanowiczKessler_submitted_draft.pdf (accessed 1 March 2016).

Janowicz, K.,Scheider, S., Pehle, T. and Hart G. (2012). Geospatial Semantics and Linked Spatiotemporal Data-Past, Present, and Future. *Semantic Web*, **3**(4): 321–32.

Jessop, M. (2008). Digital Visualization as a Scholarly Activity. *Literary and Linguistic Computing*, **23**(3): 281–93. doi:10.1093/llc/fqn016.

Murrieta–Flores, P. and Gregory, I. (2015). Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology. *Open Archaeology*, **1**(1). http://www.degruyter.com/view/j/opar.2014.1.issue-1/opar-2015-0010/opar-2015-0010.xml (accessed 1 March 2016).

Reuschel, A-K. and Hurni, L. (2011). Mapping Literature: Visualisation of Spatial Uncertainty in Fiction. *The Cartographic Journal*, **48**(4): 293–308.

Simon, R.,Isaksen, L., Barker, E. and de Soto Cañamares, P. (2015). The Pleiades Gazetteer and the Pelagios Project. In Berman, M., Mostern, R. and Southall, H. (Eds.), *Placing Names: Enriching and Integrating Gazetteers*, Indiana University Press .

Southall, H.,von Lunen, A. and Aucott, P. (2009). On the organization of geographical knowledge: Data models for gazetteers and historical GIS. *E-Science Workshops*, 2009 5th IEEE International Conference on (Oxford: IEEE), pp. 162–66.

Southall, H.,Mostern R. and Berman, M. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, **5**(2): 127–45

Tomasi, F.,Ciotti, F., Daquino, M. and Lana, M. (2015). Using Ontologies as a Faceted Browsing for Heterogeneous Cultural Heritage Collections. *1st Workshop on Intelligent Techniques At LIbraries and Archives* (IT@LIA 2015). http://italia2015.dei.unipd.it/papers/ITALIA_2015_submission_5.pdf(accessed 1 March 2016).

Van Hooland, S.,De Wilde, M., Verborgh, R., Steiner, T. and Van de Walle, R. (2013). Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections. *Literary and Linguistic Computing*. http://freeyourmetadata.org/publications/named-entity-recognition.pdf(accessed 1 March 2016).

GAMS and Cirilo: research data preservation and presentation

Martina Bürgermeister
martina.buergermeister@uni-graz.at
Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz, Austria

Zsófia Fellegi
fellegizs@pim.hu
Petőfi Literary Museum, Hungary

Gábor Palkó
palkog@pim.hu
Petőfi Literary Museum, Hungary

Gerlinde Schneider
gerlinde.schneider@uni-graz.at
Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz, Austria

Martina Scholger
martina.scholger@uni-graz.at
Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz, Austria

Elisabeth Steiner
elisabeth.steiner@uni-graz.at
Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz, Austria

Gunter Vasold
gunter.vasold@uni-graz.at
Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz, Austria

Introduction

Modern infrastructures for the management and dissemination of humanities data face various challenges. On the one hand, sustainability and availability for long-term preservation have to be guaranteed. On the other hand, flexibility and the possibility of individual data usage play a major role in this field. The FEDORA based Asset Management System GAMS (Geisteswissenschaftliches Asset Management System – Humanities' Asset Management System) and the corresponding Cirilo client, developed by the Centre for Information Modelling - Austrian Centre for Digital Humanities (ZIM-ACDH), address both demands in combining long-term preservation with a presentation and management layer. This means that all of the mentioned challenges can be solved within one infrastructure.

GAMS

GAMS is an OAIS compliant infrastructure designed for the management, publication and long-term preservation of digital resources. It is based on the Open Source software FEDORA and focuses on the long-term availability and flexible use of digital content. Since 2014 it is a certified trusted digital repository in accordance with the guidelines of the Data Seal of Approval.

FEDORA offers content models for the aggregation of the digital content, metadata and processing instructions. An example for such an aggregation could be a TEI-document with XSL transformations for various dissemination formats (HTML, PDF, different views for analysis purposes etc.), an RDF datastream that handles the description of the object's semantic relations and images with their metadata (e.g. facsimiles of the data covered in the TEI). A functional view on FEDORA's object model includes program-controlled processes based on web services e.g. XSL transformations for dynamic outputs or automatic extraction of semantic relations within a TEI document.

Cirilo

Cirilo is a java application developed for content preservation and data curation in FEDORA-based repository systems. The client operates through FEDORA's Management API (API-M) and consequently offers functionalities, which are especially suited to be used as tools for mass operations on FEDORA objects, complementing FEDORA's inbuilt administrator's client. Cirilo exploits FEDORA's object model by providing a collection of predefined content models optimized for specific primary sources like TEI, LIDO, METS/MODS, OAI-Records, HTML, PDF, BibTeX or external resources that can be accessed via a URL.

Several possibilities to semantically enrich the digital objects during the ingest process through different customizable mapping methods are offered: Dublin Core for exposing the objects in harvesting environments (like Europeana), geographical information to present data in map applications (like the DARIAH-DE Geo-Browser), or RDF statements that are stored in triplestores. Furthermore, extraction of semantic information, automatic addition of picture references to the object or resolving text data in combination with underlying ontologies are possible.

Since 2014, the Cirilo client has been available as an open-source tool including a comprehensive documentation, representing one of the Austrian contributions to DARIAH-EU. Thus, the entire infrastructure including repository and client is accessible as an archive-in-a-box solution for a broad user community.

Application

In this context, the infrastructure was adopted by the Petőfi Literary Museum as a solution for their online editions. The project DigiPhil is an online knowledge base for publishing scholarly text editions, writers' bibliographies, aggregating philological metadata and semantic annotation of these sources. At the University of Graz, GAMS hosts numerous digital editions as well as image collections and source books from various disciplines: the literary analysis of the enlightened Spectator press of the 18th century, a postcard collection with topographical and historical views on Styria from 1900 to the present or the account books of Basel from 1535 to 1610.

The workshop offers the opportunity to learn more on the underlying principles of the infrastructure and to try the mentioned functionalities of the Cirilo client in a hands-on session with a provided data sample.

Bibliography

Burghartz, S. (2015). *Jahrrechnungen der Stadt Basel 1535-1610 – digital*. http://gams.uni-graz.at/srbas (accessed 16 March 2016).

Cirilo Client. https://github.com/acdh/cirilo (accessed 16 March 2016).

Consultative Committee for Space Data Systems (2012). *Reference Model for an Open Archival Information System (OAIS), Recommended Practice*. CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012http://public.ccsds.org/publications/archive/650x0m2.pdf (accessed 16 March 2016).

DARIAH-DE. *DARIAH-DE Geo-Browser*. http://geobrowser.de.dariah.eu/ (accessed 16 March 2016)

Ertler, K., Fuchs, A., Fischer, M. and Hobisch, E. (2011-2016). *The Spectators in the international context*. http://gams.uni-graz.at/mws (accessed 16 March 2016).

Fedora Leadership Group. *FEDORA Commons*. http://www.fedora-commons.org/ (accessed 16 March 2016).

Lagoze, C., Payette, S., Shin, E. and Wilper, C. (2005). *Fedora. An Architecture for Complex Objects and their Relationships*. http://arxiv.org/ftp/cs/papers/0501/0501012.pdf (accessed 16 March 2016).

Petőfi Literary Museum.*Tudományos szövegkiadások, bibliográfiák és kutatási adatbázisokonline tudástára*. http://digiphil.hu/ (accessed 16 March 2016).

Steiner, E., Stigler, J. (2015) *GAMS and Cirilo Client. Policies, documentation and tutorial*. http://gams.uni-graz.at/doku (accessed 16 March 2016).

Stigler, J., Hofmeister, W. (2010). Edition als Interface. Möglichkeiten der Semantisierung und Kontextualisierung von domänenspezifischen Fachwissen in einem Digitalen Archiv am Beispiel der XML-basierten Augenfassung zur Hugo von Montfort-Edition. In Nutt-Kofoth, R., Plachta, B. and Woesler, W. (eds), *Editio. Internationales Jahrbuch für Editionswissenschaft*, 24/2010, pp. 39-56.

Biographical Data Workshop: modeling, sharing and analyzing people's lives

Antske Fokkens
antske.fokkens@vu.nl
VU University, Netherlands, The

Eveline Wandl-Vogt
Eveline.Wandl-Vogt@oeaw.ac.at
Austrian Academy of Sciences, Austria

Thierry Declerck
declerck@dfki.de
German Research Center for Artificial intelligence, Germany

Serge ter Braake
s.terbraake@uva.nl
University of Amsterdams, Netherlands, The

Eero Hyvönen
eero.hyvonen@aalto.fi
Aalto University, Finland

Arno Bosse
arno.bosse@history.ox.ac.uk
Oxford University, United Kingdom, The

Barbara McGillivray
arbara.mcgilli@gmail.com
Oxford University, United Kingdom, The

There is an abundance of biographical information online that begs to be analyzed with computational methods. Resources like Wiki- and DBpedia, Biographical Dictionaries, Historical Databases, Newsfeeds, Facebook and Twitter all provide information on individual's lives. 'Biographical data' is of particular interest to computer scientists, because it can be well structured and all people share common attributes such as place of birth, place of residence, parents, et cetera.

The analysis of `biographical data' with new techniques is a topic that is finding strong interest in research groups all over the world, demonstrated most recently by the first conference on Biographical Data, organized in Amsterdam in 2015. This conference brought researchers from various domains together including historians, librarians, computer scientists, data scientists, and computational linguists.

The purpose of this workshop is to take a next step in strengthening the community working with digital biographical data by exploring possibilities of turning shared interest into new international collaborations. A central theme in this next step will be connecting and linking data.

A strong international community working on biographical data and aiming for shared data representations can directly support other domains in the digital humanities: easily accessible background information on people involved in historical or contemporary events, on artists, researchers or groups of people with common professions can provide background information and therefore be of interest to digital humanities researchers working with topics beyond research on biographical data.

This workshop brings together researchers from various domains working on biographical data. In addition to sharing latest progress, it has the specific aim of initiating efforts to share (knowledge about) data and data models. The workshop directly contributes to the efforts of the DARIAH workgroup on biographical data and aims to involve new researchers in this collaboration. A call for organization will go out for the Biographical Data in a Digital World Conference in 2017 (2015 conference: http://ceur-ws.org/Vol-1399/).

The workshop consist of two main components: 1) a poster session where researchers can share their latest work on biographical data and computational analysis and 2) dedicated sessions about data and data models.

For our poster session, we explicitly invite researchers to the workshop who work with biographical data for historical research or data analysis (e.g. computational linguists, visualization experts) and are thus already very familiar with models for biographical data, but are not necessarily involved in designing them. This perspective is of great value during discussions on sharing and modeling data and can provide insights into what kind of data models are practical to work with or which links between various datasets are most valuable for research. These insights can in turn help to identify logical and practical first steps towards increasing international collaboration

Descriptions on data models and data samples will be distributed to participants in advance and studied by a panel. The panel will present their findings and support the discussion on sharing data. Direct involvement in several projects and strong relations with other international partners guarantees an interesting set of data and data models.

Digital Literary Stylistics Workshop

J. Berenike Herrmann
bherrma1@gwdg.de
Göttingen University Germany, Germany

Francesca Frontini
francesca.frontini@ilc.cnr.it
Instituto di Linguistica Computazionale “A. Zampolli”, CNR, Pisa, Italy

Marissa Gemma
marissa.gemma@aesthetics.mpg.de
Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany

The study of textual style by computational means has a long tradition, but only quite recently have such approaches been able to leave the sphere of forensics to become an instrument of legitimate literary analysis of style.

This new line of research has already yielded a considerable range of digital style studies of epochs, genres, and authors (e.g., Egbert, 2012, Hoover, 2007, Houston, 2013, Mahlberg, 2012). At the same time, questions of conceptualization (e.g., Herrmann et al., 2015) and method (e.g., Jockers, 2013, Moretti, 2013 and 2005) have quickly arisen, and it has become clear that the use of software and computational means does not necessarily imply the adherence to a strictly experimental methodological approach (cf. Ramsay, 2011).

In this workshop, we shall explore novel empirical findings as well as methodological and epistemological questions connected with the computational study of style. In particular, we intend to address topics that fall within the following domains:

(A.) The technical domain. Tools/methods of analysis (computational stylistics, corpus stylistics, literary authorship attribution, NLP, digital hermeneutics); research infrastructures and corpus building;

(B.) The empirical domain. Style variation across factors such as genre, author, gender, epoch and period; narrative perspective and characterization; non-literary registers;

(C.) The conceptual domain. Notion(s) of style; style and aesthetics; typology of style indicators; style production and reception; research design (“close”, “scalable”, and “distant reading”); epistemological status of data; issues of representativeness.

First, the workshop will feature a set of papers focused on the technical side of digital literary stylistics. **Jean-Gabriel Ganascia** will present a tool for establishing stylistic distinctiveness on the basis of syntax (“Towards a computational and syntax-based stylistics”) and **Tomoji Tabata** will evaluate a series of Eder's (2015) “Rolling Stylometry” techniques (“Experimental Stylistics: A Meta-analysis to Evaluate Rolling Stylometry”).

Next, we turn our attention to empirical findings about style. Papers in this category include **Christof Schöch**’s computationally-informed revision of Spitzer’s take on Racine (“Spitzer on Racine, digitally revisited”); **Natalie Houston**’s discussion of period-specific styles in nineteenth-century British poetry (“Towards a Computational Poetics: Some Features of Nineteenth-Century Poetic Style”); **Anne Bandry-Scubbi**’s presentation of stylistic differences between male and female authors in a corpus of novels between 1750-1830 (“Women’s Novels 1750s-1830s and the Company They Keep: A Computational Stylistic Approach”); **Jan Rybicki**’s analysis of how most frequent word usage changes with time in the oeuvre of authors of various languages and literary periods (“Authorial chronology by most frequent words: do writers’ stylometric thumbprints evolve with age?”), and **Christine Knoop**’s analysis of the distribution of rhyme and cadence schemata in German lyrical poetry between 1700-1930 (“Rhyme and Cadence Distribution in Poetry”).

Finally, the workshop will address ‘style’ conceptually within the framework of computational analysis, with papers including a discussion by **Mike Kestemont** of different ways of understanding artistic authenticity in various humanistic disciplines (“The Matter of Art: Authenticity Criticism in the Humanities”); a paper by **Sarah Allison** on what the humanistic adoption of the statistics concept of a ‘proxy’ means for stylistic research (“A Proxy for Style”); and **Fotis Jannidis** on whether the concept of “historical period” is a meaningful notion in stylometry (“Period Styles”). Expanding this conceptual discussion, **Hugh Craig** will present some possible remedies to critiques of descriptive quantitative stylistics (“Beyond Authorship”), and **Mark Algee-Hewitt** will examine how two common assumptions in digital literary stylistics (that authors are individuals with identifiable mentalities, and, in the pragmatics of authorship attribution, that these individuals can be reduced to probabilities of word frequencies) affect our notion of the author (“The Author: Between Style and Substance”).

Insights from the three domains will be drawn together in a discussion panel. Here, we will identify current topics and trends, define the horizons of “Digital Literary Stylistics” and map out the avenues of collaboration between its different branches. Our aim is to define a shared roadmap for a Special Interest Group (SIG) within the ADHO for the field of Digital Literary Stylistics. A special issue of a DH journal will comprise the papers as well as a report of this discussion.

Bibliography

Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future?. *Scientific Study of Literature*, **1**(1): 15–23 doi:10.1075/ssol.1.1.02bib.

Craig, D. H. and Kinney, A. F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press

Dalen-Oskam, K. van (2014). Epistolary voices. The case of Elisabeth Wolff and Agatha Deken. *Literary and Linguistic Computing*, **29**(3): 443–51 doi:10.1093/llc/fqu023.

Eder, M. (2015). Rolling stylometry. *Digital Scholarship in the Humanities*, **30**, first published online: 7 April 2015, doi: 10.1093/llc/gqv010.

Egbert, J. (2012). Style in nineteenth century fiction: A Multi-Dimensional analysis. *Scientific Study of Literature*, **2**(2): 167–98 doi:10.1075/ssol.2.2.01egb.

Erlin, M. (2014). *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. Boydell and Brewer.

Evert, S., et al. (2015). Towards a better understanding of Burrows's Delta in literary authorship attribution. *Fourth Workshop on Computational Linguistics for Literature, at NAACL HLT 2015*.

Ganascia, J.G., Glaudes, P. and Lungo, A. D. (2014). Automatic detection of reuses and citations in literary texts. *Literary and Linguistic Computing*, **29**(3): 412–21 doi:10.1093/llc/fqu020.

Herbelot, A. (2015). The semantics of poetry: A distributional reading. *Digital Scholarship in the Humanities*, **30**(4): 516–31.

Herrmann, J. B., Dalen-Oskam, K. van and Schöch, C. (2015). Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory*, **9**(1): 25–52 doi:10.1515/jlt-2015-0003.

Holmes, D. I. (1988). The Analysis of Literary Style--A Review. In Thoiron, P., Labbé, D. and Serant, D. (eds), *Vocabulary Structure and Lexical Richness*. Paris: Champion-Slatkine, pp. 67–76.

Hoover, D. L. (2007). Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style*, **47**: 174–203.

Houston, N. M. (2014). Toward a Computational Analysis of Victorian Poetics. *Victorian Studies*, **56**(3): 498–510 doi:10.2979/victorianstudies.56.3.498.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. (Topics in the Digital Humanities). University of Illinois Press.

Kestemont, M., et al. (2012). Cross-Genre Authorship Verification Using Unmasking. *English Studies*, **93**(3): 340–56 doi:10.1080/0013838X.2012.668793.

Klaussner, C., Nerbonne, J. and Çöltekin, Ç. (2015). Finding characteristic features in stylometric analysis. *Digital Scholarship in the Humanities*, **30**(suppl 1): 114–29.

Lynch, G. and Vogel, C. (2015). Chasing the Ghosts of Ibsen: A computational stylistic analysis of drama in translation. *SciRate* https://scirate.com/arxiv/1501.00841 (accessed 18 March 2016).

Mahlberg, M. (2012). *Corpus Stylistics and Dickens's Fiction*. 1st ed. (Routledge Advances in Corpus Linguistics). New York: Routledge http://www.routledge.com/books/details/9780415800143/.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London; New York: Verso.

Ramsay, S. (2008). Algorithmic Criticism. *Companion to Digital Literary Studies*. (Blackwell Companions to Literature and Culture). Oxford: Blackwell Publishing Professional http://www.digitalhumanities.org/companionDLS/ (accessed 24 February 2010).

Rybicki, J. and Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. *Literary and Linguistic Computing* doi:10.1093/llc/fqt027. http://llc.oxfordjournals.org/content/early/2013/05/26/llc.fqt027 (accessed 12 June 2013).

Rybicki, J., Hoover, D. and Kestemont, M. (2014). Collaborative authorship: Conrad, Ford and Rolling Delta. *Literary and Linguistic Computing*, **29**(3): 422–31 doi:10.1093/llc/fqu016.

Stajner, S. and Zampieri, M. (2013). Stylistic Changes for Temporal Text Classification. *Proceedings of the 16th International Conference on Text Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence (LNAI)*. Pilsen, Czech Republic.

A Demonstration of Multispectral Imaging

Gregory Heyworth
heyworth@olemiss.edu
The Lazarus Project, University of Mississippi, United States of America

Chet Adam Van Duzer
chet.van.duzer@gmail.com
The Lazarus Project, University of Mississippi, United States of America

Ken Boydston
ken@mega-vision.com
The Lazarus Project, University of Mississippi, United States of America; MegaVision

Michael Phelps
mphelps@emelibrary.org
Early Manuscripts Electronic Library

Roger Easton
easton@cis.rit.edu
The Lazarus Project, University of Mississippi, United States of America

Multispectral imaging is a powerful tool to recover text from manuscripts affected by fading, palimpsesting, water, fire, or overpainting. Many scholars working in the digital humanities have some acquaintance with the technology, but practical experience will help them better understand the range of imaging modalities that comprise contemporary multispectral imaging and the potential of each to advance their research. Though this workshop, scholars will be equipped to identify good candidates for the technology, and thus contribute to the recovery and preservation of our cultural heritage.

The Lazarus Project, headquartered at the University of Mississippi and directed by Gregory Heyworth, operates such a state-of-the-art multispectral imaging system that can be transported to institutions and researchers around the world so that the technology can be made available at no cost. This is part of the educational mission of the Lazarus Project – as a teaching tool for undergraduate and graduate students in the humanities and the imaging sciences – and is the reason why the capability may be offered at no cost. The system has been transported to institutions in the USA, England, Wales, France, Italy, Germany, and the Republic of Georgia to image a number of important historical objects, among which are the Vercelli Book, the Black Book of Carmarthen, and the c. 1491 world map by Henricus Martellus. This mission and capability of the Lazarus Project enables mid-sized institutions with a few manuscripts to benefit from the technology of multispectral imaging without making a sizeable investment in imaging equipment and personnel training. In addition, the cultural heritage objects are imaged at their home institution, without the difficulty of traveling to a stationary system that may be thousands of miles distant. The availability of this portable system that provides multispectral imaging free of charge makes it even more important that scholars who visit libraries and archives be able to identify good candidates for the technology.

The Lazarus Project will offer a workshop for participants in DH2016 in which the portable multispectral imaging system and the subsequent spectral image processing will be demonstrated. These demonstrations will take place in the Jagiellonian Library, and manuscripts from the library will be imaged. Thus, participants in the workshop will have the opportunity to see how multispectral imaging works, from the imaging of the object to the processing of the images and results. Each session will include a description of the equipment and process (camera, lens, LED light system, PhotoShoot software, and processing software tools), a few cycles of imaging of leaves of a manuscript, and a look at what goes into processing the images—with time for questions.

There will be four 90-minute workshops at the following times:

Monday, 11 July 2016: 9:30am - 11:00am
Monday, 11 July 2016: 2:30pm - 4:00pm
Tuesday, 12 July 2016: 9:30am - 11:00am
Tuesday, 12 July 2016: 2:30pm - 4:00pm

A maximum of 17 participants will be permitted in each session due to space constraints in the room in the Jagiellonian Library.

We look forward to seeing you at the demonstration.

Visual Network Analysis with Gephi WorkshopCollective Interpretation of DH Communities Through Twitter Networks

Mathieu Jacomy
mathieu.jacomy@sciencespo.fr
Sciences Po, médialab, France

Martin Grandjean
martin.grandjean@unil.ch
University of Lausanne, Switzerland

Paul Girard
paul.girard@sciencespo.fr
Sciences Po, médialab, France

Gephi is a free and open source network analysis software used, among other things, in social network analysis. This workshop is intended for beginners as well as confirmed users. First, we offer an introduction to the basics of Gephi, then we explore through practice the question of visual network interpretation. We provide a dataset of both Twitter hashtags and Twitter followers graphs on various topics related to the DH community.

Why network interpretation matters

Reading a network visualization can sometimes be harder than simply producing it. Once the graph has been produced, what are we supposed to look at? Nowadays, it is common to learn how to use social network analysis software such as Gephi via online tutorials, but it is often difficult to learn how to interpret the results. Based on the experience of members of the software development team and Gephi power users, we offer this workshop to help users interpret their results.

Network visualizations are exploratory rather than explanatory. As a scholar, it is important to leverage network visualization in order to find interesting insights inside your data. Exploring a network requires mobilizing external knowledge on data's context. Exploration is about generating, and not validating, hypotheses. Networks do not carry a single, clear message, and it is as fruitful for analysis as it is bad for communication. Dispelling this misunderstanding is very important if you want to fully benefit from a tool like Gephi.

The idea that a tool can analyze things for you is another misunderstanding we can help tackle. Gephi allows you to explore multiple facets of your data, but the interpretation remains to be done by the user him/herself. Users have to spend time with their data, and a workshop is a good place to introduce this data-care principle.

Once you know what to look for in a network, you will capable of finding insight but you still have to excavate some evidence. Network metrics are more capable of doing so than the visualization itself. In this workshop we will also learn to match visual features with metrics so that you can provide evidence for what you have seen. For instance, observed clusters are proven to be modularity clusters in the sense of Newman (Noack, 2009).

Workshop schedule

Part 1: Visual network analysis in a nutshell

We start the workshop with a presentation about why and how we visualize networks (Jacomy et al., 2014) and how we interpret them (Venturini et al., 2015) through a Exploratory Data Analysis method (Tukey, 1977)

Part 2: Gephi practice

In this part we explain Gephi through examples. We manipulate Gephi on screen while participants execute the same operations on their computers, using the provided datasetsAckland, R. 2013. "Web social science: Concepts, data and tools for social scientists in the digital age." SageAckland, R. 2013. "Web social science: Concepts, data and tools for social scientists in the digital age." Sage (Grandjean, 2015). The complete chain of usage will be addressed by illustrating Gephi features from basics (software installation, layout, data table…) to advanced (computing statistics, filtering, exporting…).

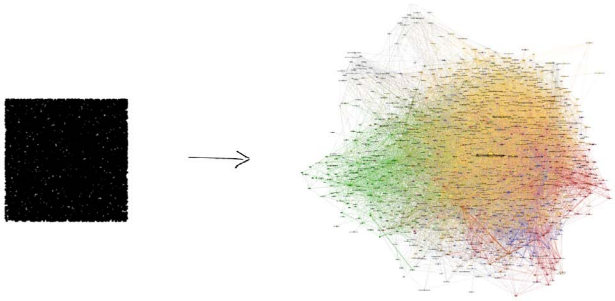


Figure 1: Participants will learn how to produce a readable Gephi map

Part 3: Guided practice

Each group makes a visualization and wraps it up in a few slides using screenshots (Girard et al., 2015). The trainers provide practical help to participants.

Part 4: Collective discussion

Each group presents its findings, and we leverage these live examples to discuss the interpretation process through networks and notably its robustness compared to the visualisation choices.

This workshop is supported by DIME-WEB part of DIME-SHS research equipment financed by the EQUIPEX program (ANR-10-EQPX-19-01).

Bibliography

Girard, P., Jacomy, M. and Plique, G. (2015). Manylines, a graph web publication platform with storytelling features Paper presented at the graph dev room, *FOSDEM*, Bruxelles, Belgique. https://archive.fosdem.org/2015/schedule/event/graph_manylines/ (accessed 14 March 2016).

Grandjean, M. (2015). GEPHI – Introduction to Network Analysis and Visualization, *Martin Grandjean*. http://www.martingrandjean.ch/gephi-introduction/ (accessed 14 March 2016).

Jacomy, M., et al. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. (Ed.) Muldoon, M. R., *PLoS ONE*, **9**(6): e98679 doi:10.1371/journal.pone.0098679.

Noack, A. (2009). Modularity clustering is force-directed layout. *Physical Review E*, **79**(2): 026102 doi:10.1103/PhysRevE.79.026102.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

Venturini, T., Jacomy, M. and Pereira, D. (2015). Visual Network Analysis: the Example of the Rio+20 Online Debate. Working paper. http://www.medialab.sciences-po.fr/wp-content/uploads/2015/06/VisualNetwork_Paper-10.pdf.

Web Communities Mapping With HypheHow to use a curation-oriented web crawler to map communities?

Mathieu Jacomy
mathieu.jacomy@sciencespo.fr
Sciences Po, France

Paul Girard
paul.girard@sciencespo.fr
Sciences Po, France

Benjamin Ooghe-Tabanou
benjamin.ooghe@sciencespo.fr
Sciences Po, France

This workshop introduces a curation-oriented web crawler called Hyphe. This software, developed with and for Social Sciences and Humanities scholars, aims at providing a method and a tool to build a research corpus from web content (web pages and HTTP links). It provides a web mining tool wrapped with a User Interface and curation features (defining web pages aggregates, filtering contents, expansion method) required by Social Sciences and Humanities scholars.

We will focus on using the web crawler and will not take the time to present web studies, digital sociology or digital methods in general. Participants should have basic knowledge of the web and already consider it as a legitimate field for scientific investigation (Ackland, 2013). Participants are encouraged to come with ideas regarding which websites would be interesting to study for their personal research agenda (a list of entry points).

Hyphe, a curation-oriented approach to web crawling for the social sciences

The web is a field of investigation for social sciences, and platform-based studies have long proven their relevance. However the generic web is rarely studied in and of itself, though it contains crucial embodiments of social actors: personal blogs, institutional websites, hobby-specific media… We realized that some sociologists see existing web crawlers as "black boxes" unsuitable for research though they are willing to study the broad web. Hyphe is a crawler which was developed with and for social scientists, with an innovative "curation-oriented" approach meant to address two of the main social science problems when working with web mining: how to build a corpus and how to delineate an actor's presence (Jacomy et al., 2016).

Workshop schedule

The workshop will first introduce Hyphe's software and methodological principles through a guided case study. The participants will be guided through their first use of Hyphe to build their own web corpus.

Part 1: Presentation of methodological approaches with a case study

We will start the workshop with a presentation of our software Hyphe and its methodological principles. It will be done through its application on a case study. We offer to map the Digital Humanities communities through the many websites used to present and organise associations, conferences, research projects, research labs… We will build such a corpus live during this first part to introduce the participants to the main concepts and practical steps one should meet when building a web corpus. The teachers will have prepared the corpus before the workshop with a series of most common use cases and issues. The subject of digital humanities is proposed first because these communities use web communication a lot, and secondly to better engage the participants with a subject they are familiar with.

Part 2: Hyphe practice

After this extensive presentation of Hyphe, participants will be invited to engage in practice themselves. Individually or as groups of two, they will be given access to their own corpus on an online version of Hyphe and will be invited to map web communities on their subject of research following Hyphe's iterative curation process:

- define the first actors web "boundaries" and start crawling them
- observe the resulting network of actors (websites)
- prospect the web for other potentially interesting actors by exploring most linked actors, filtering out irrelevant ones such as Google or Youtube
- crawl these actors' websites as well
- adjust the "boundaries" of the new actors found to better represent their social reality
- iterate over and over until obtention of a reasonably complete corpus
- visualize as a network map and take a quick look at its structural properties (clusters, density…)

We will conclude this part with a discussion on methodologies to wrap-up the workshop.

This workshop is supported by DIME-WEB part of DIME-SHS research equipment financed by the EQUIPEX program (ANR-10-EQPX-19-01).

Bibliography

Ackland, R. (2013). *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. SAGE.

Jacomy, M., et al. (2016). *Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences*. Cologne, Allemagne: AAAI https://spire.sciencespo.fr/hdl:/2441/60bemb2hsj9pb0j9bbvc7sftne.

Data mining digital libraries

Lars Gunnarsønn Johnsen
yoonsen@gmail.com
National Library of Norway, Norway

Magnus Breder Birkenes
magnus.birkenes@nb.no
National Library of Norway, Norway

Arne Martinus Lindstad
arne.lindstad@nb.no
National Library of Norway, Norway

The central theme for this workshop is data mining and the connection between metadata and data in the context of digital libraries. Digital resources and search engines raise several questions about the relationship between metadata and the data they describe. For example, what is the relationship between metadata keywords and classification categories (e.g. Dewey)? How should topics found by topic modeling algorithms be labelled? With readily available search engine technology, using document relevance based on content words, is there a need for library classification systems at all, like Dewey or UDC?

While there may be overlap between metadata and the texts described contentwise, metadata typically contain information not found within the text, such as author, geolocation and time data. In addition, subject or topic words typically consist of carefully constructed language models in the form of thesauri dedicated specifically towards specialized literary collections within different fields. The question is then how search engines may benefit from such metadata with a language model, and for what kind of library user?

In this workshop, we invite colleagues to discuss the application of various methods related to digital library resources, including the structure of the metadata itself, as well as digital book collections. Many resources are available to libraries in digital form, like journals and new book titles, while some libraries also have launched digitization programs to create digital libraries, using scanners and OCR technology.

Both the text data and the metadata of digital libraries can be scrutinized with data mining techniques, opening up the material for large-scale, quantitative analysis. This makes such collections highly relevant for Digital Humanities studies.

Background

The ongoing trend towards increased digitization in society in general poses numerous challenges at many levels, but also opens up for vast opportunities within many fields, including the library sector.

At the National Library of Norway, a mass digitization project was initiated in 2006, with the goal of digitizing the entire collection of books, newspapers, movies, radio- and television-broadcasts, music etc., in sum everything published in the public domain in Norway of all media types, i.e. the entire cultural heritage of Norway. For books, the goal is to have the entire stock digitized by 2017. Thus far, some 435.000 of 450.000-500.000 books have been digitized. When all books and newspapers have been digitized, we estimate that our Norwegian text corpus will consist of some 80 - 100 billion tokens, which is big for a rather small language like Norwegian with approximately 5 million speakers. In comparison, the Google Books corpus contains approximately 500 billion tokens for English.

The National Library cooperates with scholars of literary studies and linguistics in developing and applying methods of data mining to the digital collection. We develop services that make the content available for quantitative research, without challenging intellectual property rights. One such service is NB N-gram for Norwegian (see http://www.nb.no/sp_tjenester/beta/ngram_1/), comparable to Google Ngram Viewer for English and other languages.

Workshop leaders

Lars G. Johnsen: Research librarian at the Nation Library of Norway, PhD in linguistics. Fields of interest: semantics, grammar, philosophy of language, probability theory and applications. Email: Lars.Johnsen@nb.no, Phone: +47 23 27 61 84

Arne Martinus Lindstad: Research librarian at the National Library of Norway, PhD in linguistics. Fields of interest: corpus linguistics, language change, comparative syntax, negation. Email: arne.lindstad@nb.no, Phone: +47 23 27 62 11

Magnus Breder Birkenes: Research librarian at the National Library of Norway, PhD in linguistics. Fields of interest: corpus linguistics, history and dialectology of the Germanic languages. Email: magnus.birkenes@nb.no, Phone: +47 23 27 60 54

Call for papers (cfp)

Are you interested in automatic classification of documents and what implications this has for libraries? How may search engines (like ElasticSearch) benefit from library metadata? Do you have any experience with developing public/academic web services on top of large amounts of library data? If these questions appeal to you, this workshop may be of interest. The central theme for this workshop is data mining and the connection between metadata and data in the context of digital libraries.

We invite papers on topics such as:

- The structure of subject headings and descriptors, used in book classification (e.g. in building thesauri)
- The relationship between topic words and library classification systems
- The relationship between content words and topic words (of existing metadata, or as output from topic modeling algorithms)
- Automatic classification of digital documents
- Authorship attribution
- Development of computational services for research and the general public
- Legal issues arising with different data mining practices

Please send us an abstract of max. 500 words that is situated within the above context.

Program Committee

Oddrun Ohren (National Library of Norway)
Koenraad De Smedt (University of Bergen)
Anders Nøklestad (University of Oslo)
Elise Conradi (National Library of Norway)

Audiovisual Data And Digital Scholarship: Towards Multimodal Literacy

Martijn Kleppe
martijn.kleppe@kb.nl
National Library of the Netherlands, The Hague

Scagliola Stef
scagliola@eshcc.eur.nl
Erasmus University Rotterdam

Henderson Clara
clahende@indiana.edu
Indiana University, Bloomington

Oomen Johan
joomen@beeldengeluid.nl
Netherlands Institute for Sound and Vision

In many online platforms and websites, audio-visual data is gradually playing an equal or greater role than text. Similarly, in multiple disciplines such as anthropology, ethnomusicology, folklore, media studies, film studies, history, and English, scholars are relying more and more on audio-visual data for richer analysis of their research and for accessing information not available through textual analysis. Developing aural and visual literacy has therefore become increasingly essential for 21st century digital scholarship. While audiovisual data allows for research to be disseminated and displayed linearly, within one modality (e.g., reading a book from first to last page), it also allows for non-linear discovery and analysis, within multiple modalities (e.g., reading a webpage, browsing to a link with a sound clip, from there to a clip with film). This workshop will address both the challenges of *analyzing* audiovisual data in digital humanities scholarship, as well as the challenges of educating contemporary digital humanists on how to access, analyze, and disseminate an entire century of information generated with audiovisual media.

Other challenges to be addressed at this workshop concern issues surrounding copyright and sustainability and their impact on the dissemination and long-term access of audiovisual resources. One area of heated debate is whether certain copyright laws, which were originally instated to protect the development of new inventions, are in fact hampering the dissemination of new knowledge due to the restrictions they place on ways information may be displayed or disseminated. Some scholars contend that copyright laws will become obsolete over the next decades, while others argue that these laws will progressively restrict how scholars use audiovisual media in digital humanities research. Because of the complicated ways these copyright restrictions relate to audiovisual media, they affect films, television and sound more profoundly than digitized books with text and still images. Given these particular challenges faced by digital humanists working with audiovisual materials, a number of questions arise regarding how we might navigate these complex issues concerning copyright, sustainability and long-term storage and access. Can infrastructures essential for shepherding these digital transitions be made available for individual audiovisual digital humanities projects? Can digital humanists look to the university library as the place to support and sustain the websites, datasets and tools created by audiovisual DH research? How can digital humanists secure the massive server space needed to sustain the large-scale storage needs inherent in audiovisual DH projects? Who should oversee the recurring process of inevitable file migration and quality assurance needed for film, photo and sound formats? Are 'business-models' and their potential commercial benefit the way to go or are such arrangements overly optimistic? Are there encouraging examples of successfully sustained audiovisual DH projects that have effectively dealt with copyright issues? Will audiovisual DH scholars become increasingly dependent on philanthropic monopolistic corporations such as Google, Facebook, IBM and Microsoft to sustain their projects? What role should universities play as custodians and advocates of the knowledge produced by our audiovisual DH projects?

Workshop Overview

This full-day workshop will start with a keynote address on multimodal literacy by Dr. Claire Clivaz, Head of Digital Enhanced Learning at the Swiss Institute of Bioinformatics of Lausanne and active in #dariahTeach for which she is Head of dissemination and developer of the module Multimodal Literacies. This keynote will be followed by three sessions of paper presentations based around three themes:

- Models for training digital humanists in accessing and analyzing audiovisual collections
- Analysis and discovery models for audiovisual materials
- Copyright and sustainability

During the fourth session, workshop participants can give very short lightning talks/project pitches of max 5 minutes of ongoing work, projects or plans. Registration for this session will take place during the workshop so no submission is needed for part of the workshop. The workshop will be closed with a plenary & interactive session.

All papers will be selected by members of the Programme Committee, following a Call for Abstract which was published at https://avindhsig.wordpress.com/workshop-2016-krakowcfp/

Programme Committee

Prof. Franciska de Jong, Erasmus University Rotterdam / CLARIN ERIC (chair)

Dr. Jakob Kreuzfeld – University of Copenhagen

Prof. dr. Julia Noordegraaf – University of Amsterdam

Dr. Cord Pagenstecher – Freie Universität Berlin

Dr. Marianne Ping Huang – University of Aarhus

Dr. Willemien Sanders – Universiteit Utrecht

Dr. Khiet Truong – University of Twente

Dr. Lars Wieneke – University of Luxembourg

Organisers

This workshop is organised by the following members of the ADHO Special Interest Group AudioVisual Material in Digital Humanities (AVinDH):

Dr. Clara Henderson, Indiana University, Bloomington, USA

Dr. Martijn Kleppe, National Library of the Netherlands (KB)

Dr. Stef Scagliola, Erasmus University Rotterdam

Johan Oomen, Netherlands Institute for Sound and Vision

Big Data: Complex Systems and Text Analysis

William Kretzschmar
kretzsch@uga.edu
Department of English, University of Georgia, United States of America

Allison Burkette
burkette@olemiss.edu
Department of Modern Languages, University of Mississippi, United States of America

Jacqueline Hettel
jacqueline.hettel@asu.edu
Nexus Lab, Arizona State University, United States of America

A complex system (CS) is a system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution. The order that emerges in human language is simply the configuration of components, whether particular words, pronunciations, or constructions, that comes to occur in our communities and occasions for speech and writing. Nonlinear frequency profiles (A-curves) always emerge for linguistic features at every level of scale. Three recent books have embraced CS and developed ideas about it much more fully. Kretzschmar has demonstrated how complex systems do constitute speech in *The Linguistics of Speech* (2009), focusing on nonlinear distributions and scaling properties. Kretzschmar's *Language and Complex Systems* (2015) applies CS to a number of fields in linguistics, including a long chapter on sociolinguistics. Finally, Burkette 2016, *Language and Material Culture: Complex Systems in Human Behavior*, applies CS to both the study of language and the anthropological study of materiality. In this workshop we wish to introduce some basic ideas about complex systems, including A-curves and scaling; talk about corpus creation with either a whole population or with random sampling; and talk about quantitative methods, why "normal" statistics don't work and how to use the assumption of A-curves to talk about document identification and comparison of language in whole-to-whole or part-to-whole situations like authors or text types. A knowledge of emergent patterns in the CS of a language can cut through the problem of "noise" currently faced in NLP experiments that restrict findings to probabilities little more than chance.

We will start the workshop with a 60 minute (40 minute explanation and demonstration, 20 minute experiential learning) general introduction by Burkette to basic terms in CS such as "states" and "emergence," and also apply those principles to language in the form of nonlinear frequency distributions and scale-free networks (as from Kretzschmar 2009). The introductory section will acquaint the audience with how the operation of a CS leaves characteristic patterns in language as people use it. One feature of the introduction will be the use of a computer simulation (Kretzschmar) so that the audience can see the process in action, not just observe its end products.

We will then organize the workshop in two additional parts: 1) Hettel, CS and Corpus Creation; 2) Kretzschmar, CS and Quantitative Measurement. In each part, we will offer explanation and demonstrations for 40 minutes, and allow 20 minutes for experiential learning. Hettel will present a rationale for corpus creation using methods of random sampling. For work on language in texts, this means using either an entire population of texts (such as all the novels by one author) or, more usually, a rigorously sampled selection of texts from a population. Either an entire population of texts or a random sample is required in order to avoid undue influence from any subsection of texts, since CS distributions emerge in every subgroup of texts. Using the example of documents from the nuclear power industry, Hettel will illustrate how a random sample can be created using quotas for each variable to be investigated. Kretzschmar will discuss the problem that frequency patterns that emerge from a CS are always nonlinear, never "normal" in the sense required for use of typical Gaussian statistics. He will present a method to assess just how nonlinear a frequency profile is so that A-curves can be distinguished from normal distributions, and then will discuss how emergent frequency profiles from a CS can be usefully described and differentiated. The discussion will conclude with examples of whole-to-whole and part-to-whole comparisons that take advantage of emergent A-curve patterns.

The organizers will provide data for participants to use on their own laptops. Participants should have a spreadsheet program (Excel, or something that reads Excel files) in order to process the data.

Bibliography

Burkette, A. (2016). *Language and material culture: Complex systems in human behavior*. Amsterdam: John Benjamins.

Kretzschmar, William A., Jr. (2009). *The Linguistics of speech*. Cambridge: Cambridge University Press.

Kretzschmar, William A., Jr. (2015). *Language and complex systems*. Cambridge: Cambridge University Press.

TEI Processing Model Toolbox: Power To The Editor

Wolfgang Meier
wolfgang@existsolutions.com
eXist Solutions

Magdalena Turska
tuurma@gmail.com
eXist Solutions

Crossing the divide between encoded XML sources and tangible, published digital edition has always been a weak spot for TEI community. Recent efforts of the TEI Simple project aimed to bridge that gap with TEI Processing Model idea. TEI Processing Model Toolbox, an eXist-db based application brings the promises of TEI Simple (Rahtz et al., 2015) to life with its implementation of the processing model enhanced with an app generator, allowing to create standalone digital editions out of the box.

Publishing an edition from TEI sources so far involved tedious work on complex stylesheets and significant effort in building an application on top of it. Using the TEI Processing Model, customising the appearance of the text is all done in TEI ODD by mapping each TEI element to a limited set of well-defined behaviour functions, e.g. “paragraph” or “heading”. The TEI Processing Model specification includes a standard mapping, which can be tweaked by overwriting selected elements. Rendition styles are transparently translated into different output media types like HTML, XSL-FO, LaTeX, or ePUB. This approach easily saves thousands of lines of code for media specific stylesheets. The power of the eXist-db database and the application framework on the other hand take care of all the other core features like browsing, search and navigation.

The proposed workshop intends to introduce the concepts of the TEI Processing Model and provide a tutorial on how to use TEI Processing Model Toolbox (Meier et al, 2016) app to experiment and try out various ODDs containing processing model instructions, upload users’ own files and create a custom ODD, and, finally, generate their own, standalone edition using the App Generator. As an inspiration it will also present examples of real apps built with App Generator and other systems employing TEI Processing Model.

It is hoped that exposure to the concepts and technologies presented during the workshop will give its participants a point of exit in the task of publishing their own research data. The subject of the workshop is also strongly tied to a short paper on practical lessons from applying the TEI Processing Model that will be presented later at the DH2016, giving interested participants the opportunity to be introduced to the concept in a much more detailed manner than is possible during brief conference talk.

Bibliography

Meier, W. and Turska, M. (2016). *TEI Processing Model Toolbox Documentation*, http://showcases.exist-db.org/exist/apps/tei-simple/doc/documentation.xml?odd=documentation.odd (accessed 5 March 2015).

Rahtz, S., Mueller, M., Pytlik-Zillig, B., Turska, M. and Cummings, J. (2015). *TEI Simple Processing Model Specification*, http://htmlpreview.github.io/?https://github.com/TEIC/TEI-Simple/blob/master/tei-pm.html (accessed 5 March 2015).

From Digitization to Knowledge: Resources and Methods for Semantic Processing of Digital Works/Texts

Pierre Nugues
pierre.nugues@cs.lth.se
Lund University

Lars Borin
lars.borin@svenska.gu.se
University of Gothenburg

Nathalie Fargier
nathalie.fargier@persee.fr
Persée (Université de Lyon, ENS de Lyon, CNRS)

Richard Johansson
richard.johansson@svenska.gu.se
University of Gothenburg

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart

Sara Tonelli
satonelli@fbk.eu
Fondazione Bruno Kessler

> Internet is a revolution that will not stop "until everything is digitized", Louis Gerstner, former Chairman of IBM, quoted in *the Economist*, June 4th 1998

Description

The goal of this workshop is twofold: First, to provide a venue for researchers to describe and discuss practical methods and tools used in the construction of semantically annotated text collections, the raw material necessary to build knowledge-rich applications. We expect such tools to include lexical and semantic resources with a focus on the interlinking of concepts and entities and their integration into corpora.

A second goal is to report on the on-going development of new tools for providing access to the rich information contained in large text collections. Semantic tools and resources, notably, are reaching a quality that makes them fit for building practical applications. They include ontologies, framenets, syntactic parsers, semantic parsers, entity linkers, etc. We are interested in examples of cases that make use of such advanced tools and their evaluation in the field of digital humanities, with a specific interest on multilingual and cross-lingual aspects of semantic processing of text.

Topics of interest

- Construction and use of ontologies for text collections
- Entity nomenclatures and bridging
- Integration of lexical knowledge in text collections
- Visualization, user interfaces
- Semantic repositories: Entities and propositions
- Interlinking of concepts and entities in multilingual text
- Representing inter-textual relations
- Semantic search and information retrieval
- Tools for semantic annotation
- Timeline-based approaches such as "culturomics"
- Technical infrastructures and standards
- Quality evaluation
- Applications in digital humanities

Invited speakers

The workshop will include one, possibly two, invited speakers of international reputation.

Motivation

One of the consequences of the digital revolution is the gradual, but inexorable availability of all kinds of text in a machine-readable format. Libraries around the world scan their collections. Newspapers offer their articles on the web. Governments put their archives and laws online. A large part of what the human mind has produced: Literature, essays, encyclopedias, biographies, etc., is, or will be, accessible in a computerized form in a wide variety of languages. Within a few years, we can predict that (nearly) all text ever produced by humanity will be available in digital form: Either born digital or digitized from books, newspapers, archives, etc.

While digitization is well underway, turning the information contained in these texts into exploitable knowledge in the information society has become a major challenge as well as a major opportunity. IBM Watson and Google's knowledge graph are recent and spectacular achievements that show the significance of knowledge extraction from text. IBM Watson is a system that can answer questions in the US Jeopardy quiz show better than any human being. One of its core components is the PRISMATIC knowledge base consisting of one billion semantic propositions extracted from the English version of Wikipedia and the New York Times, while Google's knowledge graph is based on a systematic extraction of millions of entities from a variety of sources. Such technologies are defining the information age, and they have the potential to bring a much higher degree of sophistication to "distant-reading" methodology in digital humanities, enabling large-scale access to text content.

Audience

The target audience is a mix of users that would like to apply semantic processing techniques to text and researchers in this area. Users, for instance, could be interested in the extraction of entities and their association with encyclopedic text or the extraction of relations from text: date and place of birth/death, profession, etc. Researchers would describe practical techniques and algorithms that could fit the needs of the users.

Organizers

- Lars Borin, University of Gothenburg
- Nathalie Fargier, Persée (Université de Lyon)
- Richard Johansson, University of Gothenburg
- Pierre Nugues, Lund University
- Nils Reiter, Universität Stuttgart
- Sara Tonelli, Fondazione Bruno Kessler

Mining Texts with the Extracted Features Dataset

Peter Organisciak
organis2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

The HathiTrust Digital Library holds digitized copies of nearly 15 million scanned volumes from libraries around the world. These volumes are a significant resource for large-scale research: with their scale and breadth of material, a digital humanities scholar can make new inferences on history, language use, cultural trends, or even the structure of the printed word. However, access is complicated by the complexities of navigating copyright laws around the world, while use of the materials is impeded by the effort and technical demands of a researcher. To address both of these issues, the Extracted Features (EF) dataset from the HathiTrust Research Center (HTRC) provides volumes in a format that has already been cleaned, extracted, and tagged for computation use.

In this hands-on tutorial, participants will learn to use the Extracted Features dataset for text analysis alongside the HTRC Feature Reader library, equipping them to perform research on millions of publicly-accessible volumes. Through the HTRC Feature Reader, participants will be make use of popular data science tools in Python for EF dataset analysis, and will be left with demonstrative materials to repurpose in their future work.

Data

The Extracted Features (EF) dataset from the HathiTrust Research Center (Capitanu et al., 2015) provides an open and permissive download of page-level extracted information for every page of 4.8 million volumes from the HathiTrust Digital Library. A "feature" refers broadly to a quantitative measure of some property in a dataset; for example, the number of times a word appear on a page. The EF data features include part-of-speech-tagged term counts, line and sentence counts, counts of the characters occurring on the far left and far right side of a page, and inferred language probabilities. Most notably, this information is provided *for every page*. Also, headers, body, and footer have been identified and features are provided separately for each part.

In the tutorial, participants will learn the significance of each feature, such as using line counts and character information to identify the type of content on a page, or using part-of-speech tags for improving topic models based on content.

Skills

This tutorial introduces participants to introductory text analysis in Python using the Extracted Features dataset with the HTRC Feature Reader. This includes accessing term counts and other raw information, slicing within that data, visualization trends within or across books, and leading into advanced techniques like topic modeling and sentiment tagging.

The skills taught in this tutorial are underpinned by programming in Python using a popular set of data science libraries. All code examples are provided, though they are most useful if participants are comfortable in tinkering and have a familiarity with Python's basic conventions. Our intention is to make the code examples transparent enough so as to be easily modifiable by beginner users.

A participant completing the workshop will understand:

- the structure and possibilities of the HTRC Extracted Features Dataset;
- how to access the EF dataset files, both for individual and bulk use;
- how to start a Jupyter notebook, an accessible browser-based approach to data science in Python;
- the fundamentals of reading volume files, accessing metadata, and slicing and grouping token lists;
- basic visualization of EF data; and
- an advanced analytic technique modeled on recent digital humanities methods, discussed below.

The first part of the tutorial teaches the fundamental skills for working with the HTRC Feature Reader. For the final exercise of the tutorial, participants have a choice from prepared advanced exercises, which instructors will assist individually. This structure accommodates more intensive approaches in the time given, while also leaving participants with more examples for practicing their newly-acquired skills in the future.

The advanced exercises to be provided are: classification of paratext, using features suggested by Underwood (2014); visualization of sentiment trends in books as a proxy for a plot arc as previously performed by Jockers (2015), and within-book topic modeling using the inference approach presented by Organisciak et al. (2015).

Summary

The EF dataset offers a diverse and incredibly large collections of works for analysis in an easily accessible way. By providing these works as already extracted data, the EF dataset covers a large part of the text analysis workflow for researchers and is thus particularly suited for learners. This tutorial will use EF to teach text analysis through Python, using new software called the HTRC Feature Reader. By the end, students will be able to slice, group, and manipulate individual volumes for their needs, and will be familiar with techniques for modeling texts, identifying pertinent pages, and plotting trends across books.

Bibliography

Capitanu, B., Underwood, T., Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C., J. and Downie J. C. (2015). *Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain* Vol. **2**,[Dataset]. HathiTrust Research Center, doi:10.13012/j8td9v7m.

Jockers, M. L. (2015). *Revealing Sentiment and Plot Arcs with the Syuzhet Package*.

Matthew L. Jockers. Blog. http://www.matthewjockers.net/2015/02/02/syuzhet/.

Organisciak, P., Auvil, L. and Downie J. S. (2015). *Remembering books: A within-book topic mapping technique*. *Digital Humanities 2015*. Sydney, Australia.

Underwood T. (2014). *Understanding Genre in a Collection of a Million Volumes*. Interim Report. https://dx.doi.org/10.6084/m9.figshare.1281251.v1.

Reflectance Transformation Imaging (RTI) for Cultural Heritage Artefacts

Konstantinos Papadopoulos
cpapadopoulos84@gmail.com
An Foras Feasa, Maynooth University, Ireland

Introduction: Description of the Method

Reflectance Transformation Imaging (RTI) is a non-invasive/non-contact method that has its roots in the principles of raking light (light from a low angle) that has been extensively used in museums and other heritage contexts since the 1930s. RTI only requires regular photographic equipment (DSLR camera, tripod, and a portable flashlight) and freely available software developed by Hewlett Packard Labs (Malzbender et al., 2001) and the non-profit organisation Cultural Heritage Imaging (http://culturalheritageimaging.org/). Also, it can be used both indoors and outdoors and there is no limitation regarding the size or the material of the subject-matter. Using a digital camera and a light source (flash or other point light) RTI can enhance objects' subtle surface details and interactively relight them resulting in levels of information that would have been lost by using conventional photographic techniques. In addition, dynamic shading methods by a series of predefined computational algorithms can further enhance the perception of surface characteristics (Figure 1), therefore making the identification of minor details, such as flaking, scratches, and fingerprints, possible.

The method has been extensively used on a wide range of cultural objects, such as inscriptions, manuscripts, rock art, paintings, numismatics and any possible material including metal, stone, leather, paper, wax, bone, and clay (Earl et al., 2010; Kotoula and Kyranoudi, 2013; Newman, 2015; Díaz-Guardamino et al., 2015). Conservation practice has also benefitted from the capabilities of the method since it provides very high resolution and great level of detail that helps in the identification of conservation needs and the establishment of preventive conservation measures (Kotoula, 2014).

Relevance to the Digital Humanities

Many disciplines within the Humanities have a long tradition in using imaging methods for recording and investigating cultural heritage artefacts. Computational Imaging methods, such as photogrammetry, multispectral photography, and RTI provide a whole new way of extracting information from digital photographs and producing new interactive visualisations that enhance traditional modes of working with the photographed subjects. RTI is a method that has already proven its capabilities in a wide range of contexts. The use of low-cost equipment, the ease of use, the speed of capturing and processing datasets, and most importantly the results obtained, offer users many opportunities to approach artefacts in different ways beyond conventional digitisation practices. However, RTI has not been widely used in Humanities datasets, not only because text is the main source but also due to the fact that bibliographic resources about applications and technical developments of the method are relatively scattered in areas that humanists are not very familiar with, such as archaeology and computer science. In addition, such new methods of working with cultural heritage datasets bring to the front new challenges, such as the creation of interactive online repositories, different ways of sharing processes and results, community engagement (Beale and Beale, 2015), and digital preservation.

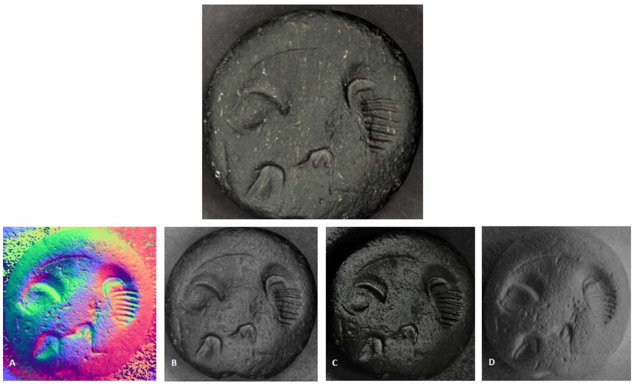


Figure 1: Reflectance Transformation Imaging on a steatite seal from Zominthos, Crete, Greece. Top: Conventional Digital Image. Bottom: Dynamic shading algorithms: A. Normal maps; B. CoefficientA5; C. Specular Enhancement; D. CoefficientA3

Method of Delivery

The tutorial will combine lecturing and hands-on practice. Participants will have the chance to learn and practice all the different stages involved in RTI - from capturing to processing and viewing as well as embedding results to webpages using the WebRTIViewer (http://vcg.isti.cnr.it/rti/webviewer.php). During the tutorial, they will capture new objects but they will also process and view pre-captured datasets. These datasets ensure that participants will work on material that has been properly captured and is of sufficient quality for all the different stages involved. Participants are also welcomed to bring artefacts that they would like to capture and process with the RTI method during the session.

Bibliography

Beale, G. and Beale, N. (2015). Community-Driven Approaches to Open Source Archaeological Imaging. In Wilson, A. T. and Edwards, B. (Eds.) *Open Source Archaeology: Ethics and Practice*. De Gruyter Open. http://www.degruyter.com/view/books/9783110440171/9783110440171-005/9783110440171-005.xml (accessed 10 March 2016).

Cultural Heritage Imaging. (2016). http://culturalheritageimaging.org/ (accessed 10 March 2016).

Díaz-Guardamino, M., García Sanjuán, L., Wheatley, D., Rodríguez Zamora, V. (2015). RTI and the study of engraved rock art: A re-examination of the Iberian south-western stelae of Setefilla and Almadén de la Plata 2 (Seville, Spain). *Digital Applications in Archaeology and Cultural Heritage*, **2**(2–3): 41-54. http://www.sciencedirect.com/science/article/pii/S2212054815300011 (accessed 10 March 2016).

Earl, G., et al. (2010). Archaeological applications of polynomial texture mapping: analysis, conservation and representation. *Journal of Archaeological Science*, **37**(8): 2040-50, http://eprints.soton.ac.uk/156253/ (accessed 10 March 2016).

Kotoula, E. (2014). Application of RTI in Museum Conservation. Archaeology in the Digital Era, Earl, G. et al. (Eds.) *Proceedings of the 40th CAA-International Conference*, **2**: 232-40, Amsterdam: Amsterdam University Press. http://pure.ltu.se/portal/files/100886592/2014_CAA2012.pdf (accessed 10 March 2016).

Kotoula, E. and Kyranoudi, M. (2013). Study of Ancient Greek and Roman Coins Using Reflectance Transformation Imaging. *E-conservation Magazine*, **25**: 74-88. http://www.academia.edu/3515894/Study_of_Ancient_Greek_and_Roman_coins_using_Reflectance_Transformation_Imaging (accessed 10 March 2016).

Malzbender, T., Gelb, D. and Wolters, H. (2001). Polynomial Texture Maps, *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*, New York, NY, USA: ACM Press, pp. 519-28. http://www.hpl.hp.com/research/ptm/papers/ptm.pdf (accessed 10 March 2016).

Newman, S. E. (2015). Applications of Reflectance Transformation Imaging (RTI) to the study of bone surface modifications. *Journal of Archaeological Science*, **53**: 536–49. http://www.sciencedirect.com/science/article/pii/S0305440314004269 (accessed 10 March 2016).

WebRTIViewer. (2016). http://vcg.isti.cnr.it/rti/webviewer.php (accessed 10 March 2016).

Building Capacity with Care: Graduate Students and DH work in the Library

Alan Gilchrist Pike
agpike@emory.edu
Emory University, United States of America

Dawn Childress
dchildress@library.ucla.edu
University of California at Los Angeles (UCLA), United States of America

Smiljana Antonijević
smiljana@smiljana.org
Pennsylvania State University, United States of America

Jim McGrath
james_mcgrath@brown.edu
Brown University, United States of America

Alex Gil
colibri.alex@gmail.com
Columbia University, United States of America

Brennan Collins
brennan@gsu.edu
Georgia State University, United States of America

Does your library or digital humanities center employ graduate students? Are you considering employing graduate students in digital scholarship work at your university? This full-day workshop will bring together institutions that do to discuss the variety of institutional arrangements for employing graduate students in digital scholarship labor. If graduate students can help build capacity for digital work on your campus, what are some best practices for structuring their employment? What are some current models in place, and what are the benefits and challenges of fellowships vs. part time employment or RAships? This workshop will present helpful practical advice on this topic, but also serve as a starting point for a broader discussion about the place of student labor in DH work.

Graduate students represent valuable members of digital humanities teams in a variety of institutional and library settings. They collaborate with scholars in labs, as members of project teams, as fellows, interns, instructors, research assistants, principal investigators, and everything in between. In her 2015 Office of Digital Humanities keynote presentation entitled "on capacity and care," Nowviskie argued that, as we continue to build individual, institutional, and even national capacity for digital scholarship in higher education, we should make an "ethic of care" the foundation upon which we work. This workshop will address how libraries and digital humanities organizations can make an ethic of care the foundation upon which their varied graduate student labor arrangements are built as they look to expand capacity within their institutions and beyond. Workshop participants will discuss the benefits, challenges, and best practices for the wide variety of institutional arrangements that result in graduate students doing DH work in libraries and DH organizations today.

Bibliography

Nowviskie, B. (2015). *On Capacity and Care. Bethany Nowviskie.* http://nowviskie.org/2015/on-capacity-and-care.

Translation Hack-a-thon!: Applying the Translation Toolkit to a Global dh+lib

Sarah Potvin
spotvin@library.tamu.edu
Texas A&M University, United States of America

Élika Ortega
elikaortega@ku.edu
University of Kansas, United States of America

Isabel Galina
igalina@unam.mx
National University of Mexico (UNAM), Mexico

Alex Gil
colibri.alex@gmail.com
Columbia University, United States of America

Daniel Paul O'Donnell
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

Patrick Williams
jpwillo3@syr.edu
Syracuse University, United States of America

Zoe Borovsky
zoe@library.ucla.edu
University of California Los Angeles, United States of America

Roxanne Shirazi
roxanneshirazi@gmail.com
City University of New York, United States of America

Zach Coble
zach.coble@nyu.edu
New York University, United States of America

Glen Worthey
gworthey@stanford.edu
Stanford University, United States of America

Efforts to forge a global digital humanities community are continually hampered by linguistic divides. Fiormonte argues that non-Anglo American DH is largely ignored by the dominant Anglo-American hegemony in the field: "But from the point of view of the scientific results, research projects, and institutional presence, Informatica Umanistica, like most of the "other" DH practiced in the world, practically doesn't exist" (Fiormonte, 2012). Rockwell asserts that lack of access to this "other" DH diminishes DH as a whole: "...it is precisely when thinkers make strange what you thought you knew that you can think about it afresh. This would be thinking-through translation. This is the message Domenico Fiormonte returns us to when arguing for multiculturalism in the digital humanities" (Rockwell, 2016).

Even when many aspects of global DH practices are not dependent on language, especially within a scholarly community that has adopted a *lingua franca*, multilingual translation is a productive avenue to explore the situatedness and locality of global DH work. It is also an indispensable basis to interlink peripheral, border, global south DH practice with mainstream and canonical DH work (Ortega, 2016). Galina poses language as a potential mode of inclusiveness, given the dominance of a few countries/institutions and English language in DH, a dominance that is reflected from academia writ large. She suggests: "… there are some indicators that there is an interest by the main DH organizations of proposing alternative models that can, if not solve, at least alleviate this phenomenon. There are two approaches: the first is making more information available in other languages and the second is making English, used as the *lingua franca*, more accessible to non-native speakers" (Galina, 2014).

How can we move beyond a monolingual DH, and promote exchange of works among linguistic communities? And how can we ensure this exchange is ongoing and sustainable? This hack-a-thon brings together practitioners from two ADHO SIGs--Global Outlook::Digital

Humanities and the Libraries and DH SIGs--and a primarily monolingual dh community project--*dh+lib*-- in an attempt to hack a solution. The half-day hack-a-thon will work on a pilot that models a translation process for a particular publication, *dh+lib*, that could be applicable to other scholarly communication vehicles and venues. The group will think through existing infrastructure, address questions around translation, labor, and design, and perform hands-on translation of works nominated by the DH and libraries community. Translation is, of course, more art than science, and any attempt to build a multilingual dh/libraries exchange must acknowledge the complexity of the undertaking. Thus the hack-a-thon builds upon previous translation exercises put into practice by both SIGs and *dh+lib*, like DH Whisperers (Ortega et al., 2015) and simultaneousbilingual publication of blog posts (Galina et al., 2015).

Despite the clear need for and benefit of broader translation in the DH community, translation is shied away from, perhaps due to a perceived inability to deal with the associated costs or develop the necessary skills. By applying the Translation Toolkit developed by GO::DH, the hack-a-thon will position translation as achievable. The Translation Toolkit gathers a catalogue of readily available tools and suggested practices to approach the sometimes daunting task of translating and preparing multilingual resources, whether at conferences, in editorial and authorial journal work, and website and resource developments. During the hack-a-thon, participants will be able to explore and put into practice the materials available in the toolkit in order to launch the translation exercises and the design of sustainable multilingual workflows at the center of the session. Further, the Translation Toolkit proposes translation and multilingual exchanges based on distributed community efforts like those put into practice by *dh+lib*.

The focus on *dh+lib* as a pilot project is notable. A community publication project, *dh+lib* was launched in 2012 by a group of librarians to enable exchange at the intersection of digital humanities and librarianship. *dh+lib* has enjoyed support from the Association of College & Research Libraries and the ADHO Libraries and DH SIG. The project includes an active site, featuring original posts, a weekly *Review* round up (using the PressForward curation tool and a multi-tiered process of editorial review), and resource pages. More than 200 volunteers and 50 authors have contributed to the site; according to GoogleAnalytics, more than 38,000 users have accessed the site. Modeled on *DHNow*, the *Review* engages volunteer editors-at-large, who develop conversance with current scholarship in digital humanities while using their Library and Information Studies expertise to bring important work by libraries to a broader audience. Editors come from a variety of professions, disciplines, and institutions (both within and outside the United States), and form a distributed community of practice connected through the collaborative hub of this participatory project. This, then, provides an ideal use case to test out the mechanisms of collaborative translation: how would a multilingual *dh+lib* function? Could this editorial system of nomination and curatorial intervention be extended to operate as a translation hub?

Building in translation as a feature of aggregation and community-based distribution will benefit the global DH community by facilitating timely cross-linguistic exposure and dialogue. Since its approval as an ADHO SIG, GO::DH has sought to “leverage the complementary strengths, interests, abilities and experiences of participants through special projects and events, profile and publicity activity, and by encouraging collaboration among individual projects, institutions, and researchers” (Global Outlook::Digital Humanities, n.d.). The collaborative hack-a-thon seeks to take full advantage of both SIGs’ pre-existing communities, the multilingual and translations initiatives previously put into practice, and the content nomination practices of *dh+lib*. The convergence of expertise will allow us to investigate approaches for sharing the labor of translation and making use of existing channels.

This hack-a-thon will prepare attendees to engage in translation work and will continue conversations around translation practices and existing workflows. It will offer participants practical and adaptable approaches to developing comfort with and practices around translation in their own institutions and endeavors. Additionally, it will provide the workshop presenters with feedback from potential users, which will help guide development of both the Translation Toolkit and a more international *dh+lib*.

We anticipate that this event, situated as it is at the beginning of DH2016, will stimulate conversations on translation and linguistic diversity that will permeate other conference events. We further expect that it will connect participants in both SIGs and serve as the basis for an ongoing collaboration throughout the year on translation workflows. We envision hosting follow-up sessions in Montreal in 2017 and Mexico City in 2018 to regroup, report progress, and continue to articulate strategies for increasing translated material among our communities, publications, and projects.

Bibliography

Fiormonte, D. (2012). Towards a Cultural Critique of the Digital Humanities. In Thaler, M. (Ed), Controversies around the Digital Humanities, *Historical Social Research / Historische Sozialforschung*, **37**(1): 59-76, Köln: Published jointly by QUANTUM [and] Zentrum für Historische Sozialforschung.

Fiormonte, D. (2014). Digital Humanities from a Global Perspective. *Laboratorio dell’ISPF*, **11**: 10.12862/ispf14L203

Galina, I. (2013). Is There Anybody Out There? Building a Global Digital Humanities Community. *Humanidades Digitales*.http://humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/.

Galina Russell, I. (2014). Geographical and Linguistic Diversity in the Digital Humanities. *Literary and Linguistic Computing*, **29**(3): 307-16.

Galina, I., Ortega, É., Priani, E. and Ricaurte, P. (2015). La RedHD y contextos latinoamericanos: auto-representación y geopolítica en las HD. *Humanidades Digitales*. http://humanidadesdigitales.net/blog/2015/02/02/la-redhd-y-contextos-latinoamericanos-auto-representacion-y-geopolitica-en-las-hd/. Also published in translation as RedHD and Latin American Contexts: Self-Representation and Geopolitics in DH. *dh+lib*. Published February 2, 2015: http://acrl.ala.org/dh/2015/02/02/redhd-and-latin-american-contexts-self-representation-and-geopolitics-in-dh/.

Gil, A. (2014). The (Digital) Library of Babel. *@elotroalex*. Published June 7, 2014: http://elotroalex.webfactional.com/digital-library-babel/.

Gil, A. and Ortega, É. (forthcoming). Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing. In Lane, R., Siemens, R. and Crompton, C. (Eds.), *Doing Digital Humanities*. London and New York: Routledge.

Global Outlook::Digital Humanities Special Interest Group. (n.d.). About. *Global Outlook::Digital Humanities*. http://www.globaloutlookdh.org/ (accessed 19 March 2016).

Global Outlook::Digital Humanities Special Interest Group. (2016). Intercultural Communication and ADHO: A GO::DH Response. Published February 12, 2016: https://docs.google.com/document/d/11b87VmblizmYeFoOUbHDOeR8A2EQ1F_Kg9jNQwUszZE/edit?usp=sharing&usp=embed_facebook.

O'Donnell, D. P., Walter, K. L., Fraistat, N. and Gil, A. (2016). Only Connect: The Globalization of the Digital Humanities." In Schreibman, S., Siemens, R.G. and Unsworth, J. (Eds.) *A New Companion to Digital Humanities*. Chichester, West Sussex, UK: John Wiley and Sons Inc., pp. 493–510.

Ortega, É. (2014). Whispering/Translating during DH2014: Five Things We Learned | Readers of Fiction. *Readers of Fiction*. Published July 21, 2014:http://lectoresdeficcion.blogs.cultureplex.ca/2014/07/21/dhwhisperer/.

Ortega, É. (2016). Crisscrossing Borders: GO::DH Regional Networks in Dialogue. *élika ortega*. Published January 13, 2016: https://elikaortega.net/2016/01/13/mla-dh-at-the-borders/.

Ortega, É., Gil, A. and O'Donnell, D. P. (2015). Psst! An informal approach to expanding the linguistic range of the Digital Humanities. *Digital Humanities 2015: Conference Abstracts*. http://dh2015.org/abstracts/xml/ORTEGA_Elika_Psst__An_informal_approach_to_expand/ORTEGA_Elika_Psst__An_informal_approach_to_expanding_th.html (accessed 16 March 2016).

Priego, E. and Gil, A. (2013). Global Perspectives: Interview with Alex Gil. *4Humanities*. Published January 11, 2013: http://4humanities.org/2013/01/interview-with-alex-gil/.

Priego, E. and O'Donnell, D. P. (2013). Bringing Diversity of Experience Together: An Interview with Daniel O'Donnell. *4Humanities*. Published May 7, 2013: http://4humanities.org/2013/05/interview-daniel-o-donnell/.

Risam, R. (2015). Across Two (Imperial) Cultures. *Roopika Risam*. Published May 31, 2015: http://roopikarisam.com/2015/05/31/across-two-imperial-cultures-2/.

Rockwell, G. (2016). Edoardo Ferrarini on the Digital Humanities in Italy. *Theoreti.ca*. Published January 28, 2016:http://theoreti.ca/?p=6010.

Terras, M. (2012). Quantifying Digital Humanities. *Melissa Terras' Blog*. Published January 20, 2012: http://melissaterras.blogspot.ca/2012/01/infographic-quanitifying-digital.html.

Minimal Computing: A Workshop

Jentery Sayers
jentery@uvic.ca
University of Victoria, Canada

Alex Gil
colibri.alex@gmail.com
Columbia University, USA

Kim Martin
kimberleymartin@gmail.com
University of Guelph, Canada

Brian Rosenblum
brianrosenblum@ku.edu
University of Kansas, USA

Tiffany Chan
tjychan@uvic.ca
University of Victoria, Canada

Scheduled for 12 July 2016, this Digital Humanities 2016 workshop will explore the practice and influence of minimal computing from both a practical and theoretical perspective. We use "minimal computing" to refer to computing done under some set of significant constraints, including constraints of hardware, software, education, network capacity, infrastructure, and power. Minimal computing is also used to capture the maintenance, refurbishing, and use of machines to do work out of necessity, along with the choice to use streamlined computing hardware, such as Raspberry Pi or Arduino.

In essence, it calls for the reduction of the technical infrastructure required to produce, disseminate, and preserve digital scholarship. Put this way, it can reduce external dependencies (such as reliance on proprietary software, network infrastructure, or complex technology stacks), help communities to assert some control over their content, and facilitate sharing and preservation. This dichotomy of choice versus necessity underscores technology that is arguably not the high-performance computing of high-income economies. By operating within this tension between choice and necessity, minimal computing brings important concepts and practices within digital humanities to the fore. In this way it is also an intellectual concept, akin to environmentalism, asking for balance between gains and costs in areas including social justice, manufacturing, waste, and labor.

The workshop will engage questions such as (but not limited to):

- What are best practices for application construction in order to maximize access, decrease obsolescence, and reduce e-waste?
- How and in what ways does experience in mid- and low-income economies inform ongoing assumptions about how research and collaboration are conducted in high-income economies?
- In terms of computing and culture, what meaningful differences emerge across economical, infrastructural, and material conditions?
- In and beyond digital humanities, what is implied by minimalist design, and to what effects on practice?
- In digital humanities and other contexts, what research is being conducted with which physical computing technologies, how, and why?
- How do the different histories of minimalism in art, design, and industry form genealogies for minimalism in computers? Or what interesting work are people currently doing with minimal computing in areas such as art, design, and experimental media?

Despite its fundamental concerns, minimal computing still lacks a cogent research agenda within digital humanities. As such, this workshop aims to bring like-minded researchers from a variety of disciplines to the same space to share work in progress and collectively articulate lines of future inquiry.

Format

The workshop will blend delivery of short papers (or thought pieces) with seminar discussion, demonstrations, and prototype testing.

9:30am-12:30pm: The first half of the workshop will consist of 8-10 presentations, together with focused discussion of the presenters' minimal computing projects. Presentations and projects will be drawn from responses to a workshop CFP circulated in March 2016.

1:30pm-4:00pm: Participants will collectively develop a research agenda for minimal computing, with all participants collaborating to identify projects, build ideas, share and test prototypes, and articulate collective interests. Where applicable, participants will demonstrate workflows and projects involving physical computing platforms such as Raspberry Pi and Arduino.

Participation

Participation in the workshop may range from presenting a paper or sharing a prototype to responding to presentations, testing prototypes, or simply observing to learn more about minimal computing practice and theory.

Working with WissKI – A Virtual Research Environment for Object Documentation and Object-Based Research

Martin Scholz
martin.scholz@fau.de
University Erlangen-Nürnberg, Germany

Dorian Merz
dorian.merz@fau.de
University Erlangen-Nürnberg, Germany

Guenther Goerz
guenther.goerz@fau.de
University Erlangen-Nürnberg, Germany

In recent years semantic technologies have become increasingly popular to represent, manage and publish data in the humanities. Virtual research environments with semantic backends are used to build complex networks, data is exposed as triples using RDF, and important vocabularies and thesauri are available as linked data. Ontologies like the CIDOC Conceptual Reference Model (CRM) are the semantic backbone of this approach and provide interoperability and data exchange beyond pure linking.

WissKI (wiss-ki.eu) is a ready-to-be-used web-based virtual research environment and publishing framework that in its core relies on Semantic Web technologies to represent the curated knowledge. The user experience for data acquisition and presentation, however, intentionally borrows from traditional modes, while the user profits from the possibilities of linked and semantically enriched data. Thus, the system enables digital humanists to produce high-quality linked data, without having to cope with technical issues of the Semantic Web and ontologies in general or the often-quoted pecularities of CIDOC CRM in particular. This is achieved by defining a mapping between traditional index card or tabular style on the one hand and graph-based linked data on the other hand. The mapping may be opaque to the users and only be managed by an (ontological) administrator. Also, mappings may be shared between systems and projects, so that best practice patterns may evolve; this actually already has happened and still happens.

By default, data may be input and displayed either as free text or as structured data via forms. Free text may be input through a graphical editor and is semantically indexed in terms of named entity recognition results, calendar date specifications, mentioned events, and also technical terms as far as appropriate authority files are available (e.g. Getty's Art and Architecture Thesaurus). Form input provides mechanisms for error reduction like spelling variants, e.g. by showing autocompletion hints that are again backed by available authorities. From the textual annotations, RDF triples may be generated and be reused as structured data. Furthermore, the system allows the upload, derivation and display of images. Other, more application-specific ways of data acquisition like mass imports or 2D/3D annotation may be included through extensions.

From the technical perspective, WissKI is based on Drupal (drupal.org). Drupal is a widely used Web Content Management System with a big and active user and developer community. It has a modular architecture and there exists a vast variety of third party extension. Being such an extension, WissKI profits from a stable core system (security updates!) and also from these community contributions, providing all sorts of functionality.

As Drupal itself, WissKI is published as open source and can be downloaded from the project web site (wiss-ki.eu) or from github.

Although WissKI in its core is domain-agnostic, it is designed to best fit the needs of object centered documentation and research as it is typical for many memory institutions, but also for research projects from art history, biodiversity, architecture, epigraphy, etc. As such it naturally goes together with the CIDOC CRM, an ontology designed for the documentation of cultural heritage. It is used by several academic and memory institutions in Germany in national and international research projects; it is used for such diverse purposes as research environment, curated collection management, virtual exhibition, or in courses and seminars.

The tutorial aims at all researchers, archivists and curators who are interested in object documentation, in particular its semantic disclosure integrating data from (database and content management systems) form-based input and plain text fields. Furthermore it addresses people interested in applications of the CIDOC CRM.

This half-day tutorial

- gives a short introduction to the (technical) approach of WissKI,
- presents current use cases and modes of use,
- shows how to install, configure, and use WissKI, and
- includes a hands-on for semantic modelling and data acquisition with WissKI and CIDOC CRM.

Digital Archiving and Storytelling in the Classroom with Omeka and CurateScape

Victoria Szabo
ves4@duke.edu
Duke University, United States of America

Hannah Jacobs
hj24@duke.edu
Duke University, United States of America

Edward Triplett
edward.triplett@duke.edu
Duke University, United States of America

Digital Archives and Exhibitions are one of the most accessible ways to bring historical and cultural materials into public circulation. This tutorial is an intensive introduction to archive development and storytelling within the Omeka content management and exhibition system (http://omeka.org/), which was developed by George Mason University's Roy Rosenzweig Center for History and New Media. This tool has been tested in a variety of academic and cultural heritage settings and was developed with extensive input and revision from the academic community. Omeka enables users to input various types of digital media content using standardized metadata structures, to organize them into Collections, and to present them in multimedia digital narratives known as Exhibitions. The platform is easy to use, and can be installed as a package in many web hosting environments. We will also demonstrate the use of the CurateScape plugin, which allows users to create location-based itineraries drawn from Omeka items optimized for mobile devices. CurateScape (http://curatescape.org), developed by the Center for Public History and Digital Humanities at Cleveland State University, can be used in a wide variety of settings. They are freely available with the only cost to the scholar being web hosting. These advantages make both tools ideal for diverse classroom settings.

Over the course of the tutorial we will introduce participants to the principles of digital archive collection development using exercises developed for the Duke University Wired! Lab for Digital Art History and Visual Culture tutorials. Content types may include digital images, audio files, 3D models, video, text facsimiles, and other source materials. These Items may be annotated with descriptive elements, locations, and other metadata relevant to search and presentation. We will explore data input formats and techniques as part of the hands-on exercises. Items in Omeka may also be organized into location-based Tours using the CurateScape Framework, a set of freely available themes and plug-ins.

In addition to providing a technical, hands-on overview of the system, we will discuss data management and digital storytelling strategies for online archives and exhibitions that take advantage of these and other tools to support their successful use in a classroom setting. We will reflect on digital pedagogy by sharing examples of successful project-based seminars that have taken advantage of Omeka based systems to scaffold course assignments around topics including discovery and remediation of analog content into digital form; integration of secondary historical and critical sources into a dynamic archive; development of taxonomies and data structures; exploration of ways to collaborate on digital media project development; recontextualization of historical objects and locations; and presentation strategies for diverse audiences in academia and in the wider public. Participants in the session will come away with a solid understanding of the features of the Omeka system, the knowledge to create their own archives, strategies for teaching with these tools, and the ability to communicate their support needs to tech professionals in their local communities.

Our plan for the session is to provide a basic Omeka installation for each participant on our shared server in the Wired! Lab, and to provide instructions to users who wish to set up their own sites on Reclaim Hosting or another hosting platform. The mobile component of the tutorial will require users be able to access wifi or local data. We will bring a few devices for users to test their projects on.

Contact Info

Victoria Szabo, Associate Research Professor, Visual and Media Studies, Duke University, ves4@duke.edu

Szabo's primary research focus is on media history and the critical and practical affordances of database-driven spatial media such as digital maps, games, virtual worlds, and mobile applications for teaching, research, and public outreach. She is especially interested in theories and practice of augmented reality experience design for digital heritage and creative expression, and has worked on location-based urban AR projects in Durham, NC and Venice, Italy, among other places. Her NC Jukebox digital audio archive project relies on the Omeka platform. She also develops art-games with the Psychasthenia Studio. She holds a PhD in English from the University of Rochester and is the Director of Duke's Information Science + Information Studies Program and the Duke Digital Humanities Initiative. Before coming to Duke she worked as an Academic Technology Specialist and Manager at Stanford University.

Hannah L. Jacobs, Multimedia Analyst, Wired! Lab, Duke University, hj24@duke.edu

Hannah holds a MA in Digital Humanities from King's College London and a BA in English/Theatre from Warren Wilson College. As Multimedia Analyst for the Wired! Lab at Duke University, she teaching web technologies, mapping, and 3D modeling tools for art history courses, as well as concepts for digital storytelling, historic mapping, and reconstruction. She teaches Omeka in a variety of course and humanities research contexts. She also consults on student-led faculty research projects and collaborates with digital humanities specialists across the university to develop workshops and resources for digital research. Hannah's research interests include digital narrative, visual storytelling, digital pedagogies, and public digital humanities.

Edward Triplett, CLIR Fellow, Duke University, edward.triplett@duke.edu

Ed recently received his PhD in the history of art and architecture from the University of Virginia. He also holds masters' degrees in 3D Modeling and Animation and in Medieval History. He is experienced with photogrammetry, virtual reality, and various interactive online presentation systems, and has instructional experience as a Visualization Specialist at UVA. As a CLIR (Council of Library and Informational Resources) postdoctoral fellow at Duke University, Ed partners with the Duke Library and the Wired! Lab to form data curation and project management plans for born digital and digitized materials. His dissertation research combined 3D modeling and GIS techniques to reconsider the architectural history of military-religious orders in medieval Iberia. His research interests include computer vision, agent-based modeling, augmented reality and data visualization.

Description of Target Audience and Number of Participants

Anyone looking for ways to implement digital storytelling and archiving tools with students. Participant limit is fifteen.

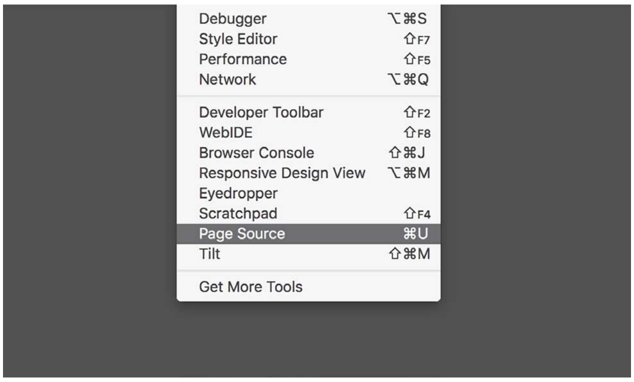
Tech Support Requirements

Participants must bring their laptops. Everyone will need internet access. Access to Omeka and CurateScape will be provided by the instructors.

View Source: Reading the Hidden Texts of the Web

Jeff Thompson
mail@jeffreythompson.org
Stevens Institute of Technology, United States of America

When discussing the web, the two layers of online experience we most often talk about are interactions and their underlying technology. We unpack the conversations and transactions happening online (anti-feminist rhetoric, how Twitter differs from Medium), and we discuss how algorithms shape those experiences (Facebook's decisions about what is in your feed, A/B testing). But at an intermediary level, an entire corpus is being written in Javascript variables and HTML comments , standardized but hidden files, and even the structure of websites themselves. This workshop will investigate these intermediary layers from a critical, exploratory point of view.



This workshop will use a hackathon-like methodology of play and exploration to maximize breadth and depth – the goal is creative and critical investigation rather than technical. After some short, fairly low-tech tutorials to introduce tools and methodologies, participants will spend the remainder of the workshop digging across the web and presenting their findings. All levels of technical expertise will be encouraged, as will ad hoc collaboration. Those with prior experience using the tools introduced are encouraged to dig deeper.

Workshop Leader

Jeff Thompson an artist, teacher, hacker, and writer whose work investigates the poetics of technology, how its functionality can provide access points for meaningful exploration, the agency of increasingly self-aware systems, and the physicalization of otherwise invisible processes. He is currently Assistant Professor and Program Director of Visual Arts and Technology at Stevens Institute of Technology.

Questions about this workshop: jeff.thompson@stevens.edu

Tech requirements

Participants should bring:

- A laptop with WiFi access
- An up-to-date version of Firefox installed (mozilla.org/firefox)
- Sublime Text text editor (free version is fine, sublimetext.com)
- If possible, please install wget (gnu.org/software/wget)
- Other text-mining/web tools of your choice

Introduction to Natural Language Processing

Lauren Tilton
lauren.tilton@yale.edu
Yale University, United States of America

Taylor Arnold
taylor.arnold@yale.edu
Yale University, United States of America

Brief Description

The application of computational tools to textual data is a growing area of inquiry in the humanities. From the culling of "Culturomics" via the 30 million document Google books collections, to the painstakingly detailed process of analyzing the text of Shakespeare's plays to ascertain their 'true' creator, a wide range of techniques and methods have been employed and developed. Text analysis in the humanities has also garnered an impressive level of interest in the mainstream media. For example, a study analyzing the relationship of a professor's gender to their teaching reviews and an overview of Franco Moretti's 'distant reading' both recently appeared in the New York Times. The Atlantic featured an historical critique of the language used in the period drama 'Mad Men', where textual analysis of the script revealed departures from the standard American English spoken in the 1960s. The majority of this work, however, relies on techniques such as n-grams and bag-of-word models. Recent developments in computational linguists, which have attempted to mimic the complex process by which humans parse and interpret language, are finding increased use within the humanities.

This workshop will introduce the basic components of modern natural language processing. Techniques include tokenization, lemmatization, part of speech tagging, and coreference detection. These will be introduced by way of examples on small snippets of text before being applied to a larger collection of short stories. Applications to stylometric analysis, document clustering, and topic detection will be briefly mentioned along the way. Our focus will be on a high-level, conceptual understanding of these techniques and the potential benefits of using them over models commonly employed for text analysis within humanities research. We will also introduce open-source software that is available for a wide range of programming languages (i.e., Java, R, Python, Ruby, Perl) and applicable for parsing an increasingly large number of natural languages (i.e., English, French, Spanish, Chinese, German, Turkish, Arabic). The workshop is based on a chapter from the instructor's book Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text (Spring, 2015).

Instructors

Taylor Arnold, is currently a lecturer in the department of statistics at Yale and senior scientist at ATandT Labs. His research focuses on the analysis of large, complex datasets and the resulting computational challenges. A particular area of focus is the sparse representation of highly structured objects such as text corpora and digital images. He is the technical co-director of the NEH funded project Photogrammar. Together with Lauren Tilton, he is the co-author of the text *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*.

Lauren Tilton, is a doctoral candidate in American Studies at Yale University. She is the Co-Director of Photogrammar (photogrammar.yale.edu) and Participatory Media (http://participatorymediaproject.org/). Research interests include 20th century U.S. history and visual culture as well as digital and public humanities. She is the co-author with Taylor Arnold of *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*. She will be joining the faculty at the University of Richmond as a Visiting Assistant Professor of Digital Humanities in Fall 2016.

Target Audience and Size

This workshop is accessible to participants from all backgrounds.

Brief Outline of Workshop

- Introduction to NLP
- Tokenization and Sentence Splitting
- Lemmatization
- Part of Speech Tagging
- Dependencies
- Named Entity Recognition
- Coreference resolution
- Overview and comparison of current software
- Stanford CoreNLPApache OpenNLPspaCy.io

Music Information Retrieval Algorithms for Oral History Collections

Sharon Webb
sharon.webb@sussex.ac.uk
Sussex Humanities Lab, University of Sussex

Chris Kiefer
c.kiefer@sussex.ac.uk
Sussex Humanities Lab, University of Sussex

Ben Jackson
b.j.c.jackson@sussex.ac.uk
Sussex Humanities Lab, University of Sussex

Alice Eldridge
alicee@sussex.ac.uk
Sussex Humanities Lab, University of Sussex

James Baker
james.baker@sussex.ac.uk
Sussex Humanities Lab, University of Sussex

Digital humanities, as a largely text based domain, often treats audio files as texts, retrieving semantic information in order to categorise, sort, and discover audio. This workshop will treat audio as audio. Taking oral history collections from the University of Sussex Archive of Resistance Testimony as a test case, participants will be lead through the use of Music Information Retrieval (MIR) approaches to categorise, sort, and support their discovery of an audio collection. Participants will also be supported in planning the extension of these approaches to explore audio collections that they know or work with.

All instructors for this workshop are from the new inter-disciplinary Sussex Humanities Lab at the University of Sussex. The workshop combines the diverse interests of the instructors to focus on the application of music technology to history and stems from this inherently collaborative environment.

Oral history best practice publications and resources often focus on the application and use of digital methods and tools to create, store and manage audio, audio-visual, and subsequent text files. They recommend, for example, standards for file formats, metadata and text encoding, software for audio to text conversion, and database and content management systems. However, while a number of projects provide innovative and useful tools that challenge the privilege of the text (i.e. Oral History Metadata Synchroniser), the majority of projects rely on the ability to encode an oral history interview to carry out further analysis using digital tools and methods. The analysis, therefore, is based on the text surrogate rather than the original audio source, but as Alessandro Portellii states this focus denies the 'orality of the oral source'.

Text encodings or transcripts of oral history interviews have their obvious advantages - they are easier to anonymise, distribute and store, and we have established techniques for text analysis. However, there are indications within the community that the privilege of this text based approach should be questioned given the ever increasing possibility for computational analysis of audio. This is evident in the Oral History Society's recent call for papers which remarks that the 'auditory dimension of oral history was for decades notoriously underused'.

In light of this loss of context and information this workshop will explore the original sources for the richer datasets which they afford. The field of MIR provides this opportunity. MIR draws from digital audio signal processing, pattern recognition, psychology of perception, software system design, and machine learning to develop algorithms that enable computers to 'listen' to and abstract high-level musical information from low-level audio data. Just as human listeners can recognize pitch, tempo, chords, genre, song structure etc, MIR algorithms are capable of recognizing and extracting this information, enabling systems to perform extensive sorting, searching, music recommendation, metadata generation, transcription on large data sets. Deployed initially in musicology research and more recently for automatic recommender systems, the research potential for MIR tools in non-musical audio data mining is being recognised (e.g. analysing bio-acoustic data for ecological purposes) but yet to be fully explored in the humanities.

The aim of this workshop is to help the digital humanities community explore the possibilities of MIR in a practical and methodological fashion within context of oral history collections. Our participatory design approach enables development for a broader scope of problems and support participants to challenge current methodologies for oral history analysis. We seek to apply an objective analysis based on the content of the source material in its entirety and complement or challenge human and text based analysis with computational methods. By applying MIR algorithms to a field that traditionally privileges text we hope to add new understandings and interpretations to rich audio resources.

Bibliography

Grele, R. (2007). Oral History as Evidence. In Thomas L. Carlton, T.L., Lois E. Myers, L.E., & Sharpless, R. (eds) *History of Oral History: Foundations and Methodology*. Plymouth: The Rowman & Littlefield Publishing Group, p. 33-94.

Index